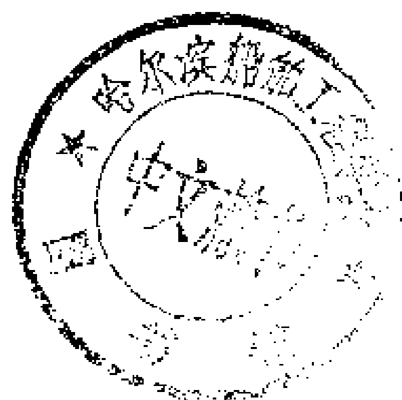
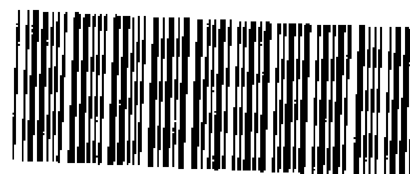
非参数统计讲义

孙山泽　编著





北京大学出版社
·北　京·

图书在版编目(CIP)数据

非参数统计讲义/孙山泽编著. —北京：北京大学出版社，2000.4

ISBN 7-301-04466-6

Ⅰ. 非… Ⅱ. 孙… Ⅲ. 非参数统计-高等学校-教材 Ⅳ. O212.7

书　　　名：非参数统计讲义
著作责任者：孙山泽　编著
责 任 编 辑：刘　勇
标 准 书 号：ISBN 7-301-04466-6/O·457
出　版　者：北京大学出版社
地　　　址：北京市海淀区中关村北京大学校内　100871
网　　　址：http://cbs.pku.edu.cn/cbs.htm
电　　　话：出版部 62752015　发行部 62559712　编辑部 62752032
电 子 信 箱：zpup@pup.pku.edu.cn
排　印　者：北京大学印刷厂
发　行　者：北京大学出版社
经　销　者：新华书店
850 毫米×1168 毫米　32 开本　7.5 印张　190 千字
2000 年 4 月第一版　　2000 年 4 月第一次印刷
印　　　数：0001—4000 册
定　　　价：12.50 元

内 容 简 介

本书是非参数统计方法的基础课教材. 书中介绍了非参数统计的基本概念和理论,着重于在统计各应用领域常用的方法,以及非参数统计推断的一般处理技术和原则. 本书系统地阐述了有广泛应用的 U 统计量、线性秩统计量及样本次序统计量线性组合等的基本性质,研究了它们在假设检验和估计等统计推断中的应用. 全书共分九章. 内容包括: U 统计量,线性秩统计量,非参数假设检验功效的研究,样本次序统计量构成的非参数估计,Hodges Lehmann 估计,M 估计等.

具有高等数学及概率统计基本知识的读者即可读懂本书的主要内容. 本书可作为大学本科高年级学生或硕士研究生学习非参数统计方法的教材,亦可作为从事统计及其他领域实际工作者查阅非参数统计方法的参考书.

序　言

本书是在先前编写的一本非参数统计油印讲义的基础上整理而成。该讲义曾在北京大学概率统计系作为“非参数统计”课程的教材使用多年，这次成书在内容及编排上都作了较大的修改，力图更能反映非参数统计领域的基础理论和一般研究方法。统计学的基本任务是利用观测的样本去推断总体的一些性质，推断过程中经常要对研究的总体做一些假定，基础数理统计的许多方法对总体的分布假定了一个参数模型，未知的是模型中的参数，问题是估计这些未知参数或对它们作某种假设检验。非参数统计方法一般对研究的总体不作具体的模型假定，只有一些定性的描述，在这样比较弱的假定下，对总体的一些未知特征进行种种统计推断。这是一个范围非常宽广的研究领域，本书涉及的仅是这一领域最基础的一些内容。书中介绍了非参数统计的基本概念和理论，着重于在统计各应用领域常用的方法，以及非参数统计推断的一般处理技术和原则，系统地阐述了有广泛应用的 U 统计量、线性秩统计量及次序统计量线性组合等的基本性质，研究了它们在假设检验和估计等统计推断中的应用。阅读本书需要有微积分、概率论、数理统计等先行课程的知识。

全书前五章介绍各种假设检验问题的模型、检验方法，以及非参数检验如何考虑检验的功效等，重点讨论 U 统计量和线性秩统计量。第六、七、八章涉及非参数点估计问题，讨论了由样本次序统计量构成的一些估计、Hodges-Lehmann 估计、以及 M 估计等一些稳健估计。第九章介绍了一些多总体非参数统计问题。

非参数统计推断的研究方法有一定的特点，这与对总体的假定有关，由于对总体分布假定很少，很难根据假定由逻辑推理寻求

最优推断，一般都是先由直观确定一个可行的推断方法，然后再论证方法的优劣，而且通常对各种方法进行相对比较，确定相对的优劣。非参数统计较强地依赖一定的统计直观能力，这种能力的培养对一个数理统计工作者应该是非常重要的。阅读本书除了获得各种统计推断的结论、统计量的基本性质、证明的逻辑推导等知识外，还应注意对各种问题的统计直观判断。

北京大学数学科学学院概率统计系郑忠国教授审阅了全书，提出了许多中肯的修改意见，北京大学出版社刘勇先生为本书的出版付出了辛勤的劳动，对书中的许多叙述提出了许多非常具体的建议，在此一并表示衷心的感谢。

本书虽曾作为教材使用多年，但限于作者的水平，书中可能仍有不少缺点和错误，欢迎各地专家和读者批评指正。

孙山泽

2000年1月15日

于北京大学燕北园

目　录

第一章 计数统计量和秩统计量

§1.1 适应任意分布的(distribution-free)统计量

统计学的基本问题是利用观测的样本推断总体的一些性质. 在统计推断中我们对研究的总体经常要做一些假定,例如在基础的数理统计教程中常假定总体是正态分布的. 著名的 t 统计量、χ^2 统计量、F 统计量均基于这一假定. 在经典的数理统计中假定样本观测来自具有某些参数的已知的分布族,问题是估计这些未知的参数或对它们作某种假设检验. 这样的假定是比较强的. 将在这些参数模型下发展的统计方法用于实际问题时,当实际的总体与假定有些差距时,常引起无法预料的错误. 这就促使人们研究在总体分布比较弱的假定下,如何进行统计推断. 非参数统计正是这样一个领域. 这是一个很广阔的研究领域,本书涉及的仅是最基本的一些内容,也就是通常非参数统计课程讲授的内容——秩统计量、U 统计量、次序统计量线性组合等的性质和它们的应用.

我们首先叙述一个在非参数统计中常用的概念——适应任意分布的统计量.

定义 1.1 设随机变量 $X_1, X_2, \cdots, X_n$ 是来自总体 $F(x)$ 的样本,一切可能的 $F(x)$ 组成分布类 $\mathscr{F}$. 如果统计量 $T(X_1, X_2, \cdots, X_n)$ 对任意的 $F \in \mathscr{F}$ 均有相同的分布,则称 T 关于 $\mathscr{F}$ 是**适应任意分布的**(d-free).

这一概念在数理统计课程中实际上也是有的,在做假设检验时,我们总要找一个对原假设 H_0 包含的分布类是适应任意分布的统计量.

例 1 设随机变量 $X_1, X_2, \cdots, X_n$ 相互独立同分布，分布为 $N(\mu_0, \sigma^2)$，参数 μ_0 已知，σ^2 未知. 分布类

$$\mathscr{F} = \{N(\mu_0, \sigma^2) \mid \sigma > 0\},$$

取统计量

$$T(X_1, X_2, \cdots, X_n) = \sqrt{n}(\overline{X} - \mu_0)/S,$$

其中 $\overline{X} = \sum_{i=1}^{n} X_i/n$，$S = \left[\sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)\right]^{\frac{1}{2}}$. 则统计量 T 关于分布类 $\mathscr{F}$ 是适应任意分布的. 对一切 $\sigma > 0$，T 的分布均为 $(n-1)$ 个自由度的 t 分布.

例 2 设随机变量 $X_1, X_2, \cdots, X_n$ 相互独立同分布，分布为 $F\left(\frac{x}{\eta}\right)$，参数 η 未知. 分布类

$$\mathscr{F} = \left\{F\left(\frac{x}{\eta}\right) \mid F(t) \text{ 为一固定分布}, 0 < \eta < \infty\right\},$$

取统计量

$$T = \frac{X_{(1)}}{X_{(n)} - X_{(1)}} = \frac{X_{(1)}/\eta}{(X_{(n)} - X_{(1)})/\eta},$$

其中 $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$ 是 $X_1, X_2, \cdots, X_n$ 的次序统计量. 相应的 $X_{(1)}/\eta, X_{(2)}/\eta, \cdots, X_{(n)}/\eta$ 是 $X_1/\eta, X_2/\eta, \cdots, X_n/\eta$ 的次序统计量，而 X_i/η 有分布 $F(x)$. 所以统计量 T 关于分布类 $\mathscr{F}$ 是适应任意分布的.

非参数统计中涉及的分布类往往是例 2 那样没有明确分布形式的. 当考虑大样本统计推断时，还会考虑渐近适应任意分布的统计量.

定义 1.2 设随机变量 $X_1, X_2, \cdots$ 是来自总体 $F(x)$ 的样本. 一切可能的总体分布 $F(x)$ 组成分布类 $\mathscr{F}$. 如果统计量 $\{T_n(X_1, \cdots, X_n), n = 1, 2, \cdots\}$ 对任意的 $F \in \mathscr{F}$ 均有相同的极限分布，则称 $\{T_n\}$ 关于 $\mathscr{F}$ 是**渐近适应任意分布的**.

例 3 设随机变量 $X_1, X_2, \cdots$ 相互独立同分布，分布为 $F(x)$.

分布类

$$\mathscr{F} = \{\text{期望为 } \mu, \text{方差为 } \sigma^2 > 0 \text{ 的分布}\},$$

取统计量

$$S_n = \sqrt{n}\,(\overline{X} - \mu)/\sigma, \quad \overline{X} = \sum_{i=1}^{n} X_i/n,$$

则 S_n 关于 $\mathscr{F}$ 是渐近适应任意分布的统计量，其极限分布均为 $N(0,1)$.

§1.2　计数统计量

定义 2.1　设 X 是一个随机变量. 对一给定的实数 θ_0，定义随机变量

$$\Psi = \psi(X - \theta_0),$$

其中

$$\psi(t) = \begin{cases} 1, & \text{当 } t > 0 \text{ 时}, \\ 0, & \text{当 } t \leqslant 0 \text{ 时}. \end{cases} \tag{1.2.1}$$

随机变量 Ψ 称为 X 按 θ_0 分段的**计数统计量**.

定理 2.1　设随机变量 $X_1, X_2, \cdots, X_n$ 相互独立. X_i 有分布 $F_i(x)$ $(i=1,2,\cdots,n)$. 又对 $i=1,2,\cdots,n$ 均有 $F_i(\theta_0)=p_0(0<p_0<1)$，则 $\Psi_i=\psi(X_i-\theta_0)$ $(i=1,2,\cdots,n)$是相互独立同分布的随机变量，其共同分布是参数为$(1-p_0)$的两点分布.

证明　显然.

由定理 2.1 可知，任一由 $\Psi_1, \Psi_2, \cdots, \Psi_n$ 构成的统计量 $S(\Psi_1, \Psi_2, \cdots, \Psi_n)$关于一切 p_0 分位点为 θ_0(θ_0 是连续点)的分布类是适应任意分布的. 这样的统计量统称为**计数统计量**. 最常用的一个计数统计量为

$$B = \sum_{i=1}^{n} \Psi_i. \tag{1.2.2}$$

例 1　**符号检验**. 设随机变量 $X_1, \cdots, X_n$ 相互独立同分布，分

布为 $F(x)$，$F(x)$在 $x=0$ 连续. 假设检验问题

$$H_0:\ F(0)=\frac{1}{2},\qquad H_1:\ F(0)\neq\frac{1}{2}.$$

检验统计量可取 $B=\sum_{i=1}^{n}\Psi_i=\sum_{i=1}^{n}\psi(X_i)$，即为 $X_1,\cdots,X_n$ 中取正号的个数. 在 H_0 下，B 的分布是参数为 n 和$\frac{1}{2}$的二项分布 $b\left(n,\frac{1}{2}\right)$. 可确定 c_1,c_2 使

$$P\{B\leqslant c_1\}\leqslant\alpha/2,\quad P\{B\geqslant c_2\}\leqslant\alpha/2.$$

否定域取为$\{0,1,\cdots,c_1\}\cup\{c_2,c_2+1,\cdots,n\}$. 这一检验的真实水平为 $\alpha'=P\{B\leqslant c_1\}+P\{B\geqslant c_2\}$.（今后我们认为 α 即为真实水平，不讨论随机化的检验.）

符号检验在实际工作中常用于对比试验. 若 X 是未加某一处理的记录，而 Y 是施加某一处理的记录，要判断有无处理效应，常研究随机变量 $Y-X$. 若无处理效应，X 和 Y 的地位是相同的，$Y-X$ 与 $X-Y$ 有相同分布. $Y-X$ 的分布关于 0 点对称，因而有$F(0)=1/2$. 以 $Y_1-X_1,\cdots,Y_n-X_n$ 为样本，检验 H_0: $F(0)=1/2$，可以用符号检验，若否定 H_0，则判断有处理效应.

例 2 感官检查中的两点比较法. 有两种产品 A 和 B，每次取一个 A 和一个 B 同时或相继出示给检验者，要求确定出哪个好或哪个刺激性强等等. 具体操作是同一人作 n 次或 n 个人一人一次. 由所得的 n 个数据判定两种产品之间是否有差异或判定检验者是否有识别这种差异的能力.

这是典型的对比试验. 当产品完全没有差异或检查者胡乱判定时，A 和 B 是随机对等的，选中哪一个的概率都是 1/2. 在 n 次判定中选中 A(或 B)的次数即为符号检验统计量 $B=\sum_{i=1}^{n}\Psi_i$，它遵从二项分布 $b(n,1/2)$. 给定检验水平 α，可确定相应的单侧或双侧检验的否定域. 当由样本计算的统计量 B 的值落入否定域时，则判定两种产品有显著差异或检查者有识别差异的能力.

感官检查中常用的一对二比较法也是一种稍作变化的符号检验. 此法用于识别两种产品的差异. 取一个 A 作标准品, 另给一个 A 和一个 B, 判别哪个与标准品相同. 当 A 与 B 没有差异时, 判定 A 试品或 B 试品与标准品相同的概率都是 1/2.

区间估计常可利用相应的假设检验的统计量来构造. 利用符号检验统计量可构造总体中位数的置信区间.

例 3 设随机变量 $X_1, X_2, \cdots, X_n$ 相互独立同分布, 分布函数 $F(x)$ 连续, 中位数记为 θ, 则

$$\Psi_i = \psi(X_i - \theta), \quad i = 1, 2, \cdots, n$$

是独立同分布的随机变量, 其共同分布为取值 0,1 的概率均为1/2 的两点分布. $B = \sum_{i=1}^{n} \Psi_i$ 有二项分布 $b\left(n, \frac{1}{2}\right)$. 可确定一正整数 c $(c \leqslant n)$, 使

$$P\{B \leqslant c\} \geqslant 1 - \alpha,$$

则 θ 的集合

$$\left\{\theta \middle| B = \sum_{i=1}^{n} \psi(X_i - \theta) \leqslant c \right\}$$

即为 θ 的置信水平为 $1-\alpha$ 的单侧置信区间. 上述集合可如下解得关于 θ 的显式:

$$\begin{aligned}
&\left\{ \sum_{i=1}^{n} \psi(X_i - \theta) \leqslant c \right\} \\
&\quad \Longleftrightarrow \{X_1 - \theta, \cdots, X_n - \theta \text{ 中至少有 } n - c \text{ 个为负值}\} \\
&\quad \Longleftrightarrow \{X_{(n-c)} - \theta \leqslant 0\} \\
&\quad \Longleftrightarrow \{X_{(n-c)} \leqslant \theta\},
\end{aligned}$$

其中 $X_{(n-c)}$ 为 $X_1, \cdots, X_n$ 的第 $n-c$ 个次序统计量.

§1.3 秩 统 计 量

定义 3.1 设 $Z_1, Z_2, \cdots, Z_N$ 是来自连续分布 $F(z)$ 的简单随

机样本，$Z_{(1)} \leqslant Z_{(2)} \leqslant \cdots \leqslant Z_{(N)}$为其次序统计量. 定义随机变量

$$R_i = r, \quad \text{当 } Z_i = Z_{(r)}, \quad i = 1,2,\cdots,N. \qquad (1.3.1)$$

当 R_i 是唯一确定时，称样本观测值 Z_i **有秩** $\boldsymbol{R_i}(i=1,2,\cdots,N)$.（由于 $F(z)$连续，因而 R_i 不唯一确定的概率为 0).

R_i 是第 i 个样本单元 Z_i 在样本次序统计量$(Z_{(1)}, Z_{(2)}, \cdots, Z_{(N)})$中的位置.

定理 3.1 记 $R=(R_1,R_2,\cdots,R_N)$，R_i 的定义见定义 3.1，集合

$$\mathscr{R} = \{(r_1,\cdots,r_N) \mid (r_1,\cdots,r_N) \text{ 是}(1,\cdots,N) \text{ 的一个排列}\}, \qquad (1.3.2)$$

则 R 在 $\mathscr{R}$ 上均匀分布.

证明 易知 R 仅在 $\mathscr{R}$ 上取值. 对任意一个 $r=(r_1,\cdots,r_N) \in \mathscr{R}$，

$$\begin{aligned} P\{R = r\} &= P\{(R_1,\cdots,R_N) = (r_1,\cdots,r_N)\} \\ &= P\{(Z_1,\cdots,Z_N) = (Z_{(r_1)},\cdots,Z_{(r_N)})\} \\ &= P\{(Z_{d_1},\cdots,Z_{d_N}) = (Z_{(1)},\cdots,Z_{(N)})\} \\ &= P\{Z_{d_1} < \cdots < Z_{d_N}\}, \end{aligned}$$

其中 $d_i=k$，当 $r_k=i$ 时，即 d_i $(i=1,\cdots,N)$是 i 在排列 r 中的位置. 又由于

$$(Z_1,\cdots,Z_N) \xlongequal{d} (Z_{d_1},\cdots,Z_{d_N})$$

（记号$\xlongequal{d}$ 表示等式两端的随机向量有相同的分布），所以

$$P\{R = r\} = P\{Z_1 < \cdots < Z_N\}$$

对任意 $r \in \mathscr{R}$，上式均成立，所以对任意 r，这个概率均相等. 而全部这样的事件互不相容且它们的和是必然事件，故对任意 $r \in \mathscr{R}$，有

$$P\{R = r\} = 1/N!.$$

证毕.

定理 3.2 $R=(R_1,\cdots,R_N)$的边缘分布也是均匀分布. 特别一维边缘分布有

$$P\{R_i=r\}=\begin{cases}\dfrac{1}{N}, & \text{当 } r=1,\cdots,N \text{ 时},\\ 0, & \text{其他};\end{cases} \tag{1.3.3}$$

二维边缘分布当 $i\neq j$ 时,有

$$P\{R_i=r,R_j=s\}=\begin{cases}\dfrac{1}{N(N-1)}, & r\neq s,\ r,s=1,\cdots,N,\\ 0, & \text{其他}.\end{cases} \tag{1.3.4}$$

证明 当 $r\neq 1,\cdots,N$ 时,显然有 $P\{R_i=r\}=0$. 当 $r=1,\cdots,N$ 时,因为

$$(Z_1,Z_2,\cdots,Z_N)\overset{d}{=\!=}(Z_2,Z_1,\cdots,Z_N),$$

从而 $R_1\overset{d}{=\!=}R_2$. 类似地可证 $R_1\overset{d}{=\!=}R_i$ $(i=2,3,\cdots,N)$. 故有

$$P\{R_1=r\}=P\{R_2=r\}=\cdots=P\{R_N=r\},$$

且有

$$\{R_i=r\}\cap\{R_j=r\}=\varnothing,\quad \text{当 } i\neq j \text{ 时};$$

$$\sum_{i=1}^{N}P\{R_i=r\}=1.$$

所以

$$P\{R_i=r\}=\frac{1}{N}.$$

类似地可证二维边缘分布和高维边缘分布是均匀分布. 证毕.

定理 3.3 对秩统计量 $R=(R_1,\cdots,R_N)$,有

$$E(R_i)=\frac{N+1}{2},\quad i=1,\cdots,N, \tag{1.3.5}$$

$$\operatorname{var}(R_i)=\frac{(N+1)(N-1)}{12},\quad i=1,\cdots,N, \tag{1.3.6}$$

$$\operatorname{cov}(R_i,R_j)=-\frac{(N+1)}{12},\quad i\neq j,\ i,j=1,\cdots,N. \tag{1.3.7}$$

证明 由定理 3.2 知，

$$E(R_i)=\sum_{r=1}^{N} r\cdot\frac{1}{N}=\frac{1}{N}\cdot\frac{N(N+1)}{2}=\frac{N+1}{2},$$

$$\begin{aligned}\operatorname{var}(R_i)&=E(R_i^2)-(E(R_i))^2=\frac{1}{N}\sum_{r=1}^{N} r^2-\left(\frac{N+1}{2}\right)^2\\&=\frac{1}{N}\cdot\frac{N(N+1)(2N+1)}{6}-\left(\frac{N+1}{2}\right)^2\\&=\frac{(N+1)(N-1)}{12}.\end{aligned}$$

由于有

$$\begin{aligned}0=\left[\sum_{r=1}^{N}\left(r-\frac{N+1}{2}\right)\right]^2&=\sum_{r=1}^{N}\left(r-\frac{N+1}{2}\right)^2\\&+\sum_{r\neq s}\left(r-\frac{N+1}{2}\right)\left(s-\frac{N+1}{2}\right),\end{aligned}$$

故

$$\begin{aligned}\operatorname{cov}(R_i,R_j)&=\frac{1}{N(N-1)}\sum_{r\neq s}\left(r-\frac{N+1}{2}\right)\left(s-\frac{N+1}{2}\right)\\&=\frac{-1}{N(N-1)}\sum_{r=1}^{N}\left(r-\frac{N+1}{2}\right)^2\\&=\frac{-1}{N(N-1)}\cdot\frac{N(N+1)(N-1)}{12}\\&=-\frac{N+1}{12}.\end{aligned}$$

证毕.

由定理 3.1 可知：仅依赖 R 的统计量 $S(R)$关于连续分布构成的分布类是适应任意分布的.

秩统计量的一个重要情形是 Wilcoxon 统计量. 其定义与理论背景如下：

有两个总体. 一个总体的样本为 $X_1,\cdots,X_m$，相互独立同分布，分布为 $F(x)$；另一个总体的样本为 $Y_1,\cdots,Y_n$，相互独立同分布，分布为 $G(x)$，$F(x)$，$G(x)$连续. 问随机变量 Y 是否大于随机

变量 X，即检验

H_0: $F(x) \equiv G(x)$,

H_1: $F(x) \geqslant G(x)$，且有某些点不等号成立.

将 $X_1, \cdots, X_m; Y_1, \cdots, Y_n$ 共 $m+n$ 个随机变量一起排序，产生对应的秩

$$R = (Q_1, \cdots, Q_m;\ R_1, \cdots, R_n),$$

Wilcoxon 统计量为

$$W = \sum_{i=1}^{n} R_i, \tag{1.3.8}$$

即 $Y_1, \cdots, Y_n$ 在混合样本中的秩的和，通常称为 Wilcoxon 秩和统计量. 当备择假设 H_1 为真时，W 的值应较大. 从而对上述检验问题可用统计量 W 构成否定域. 否定域的形式为

$$W \geqslant \omega(\alpha; m, n),$$

$\omega(\alpha; m, n)$ 是由 W 在 H_0 下的分布按检验水平 α 确定的临界值.

例 1 在农业、工业等实践中，为了要确定某种新的处理方法是否有明显的效益，常常做对照实验. 如为了确定某种新肥料是否增加作物产量，在 m 个实验地块不施用该肥料作为对照组，在 n 个实验地块施用该肥料作为处理组. 设不施用肥料条件下，作物产量的分布为 $F(x)$，施用肥料条件下，可获得增产效应 Δ，则此时作物产量的分布为 $F(x-\Delta)$. 对照实验得到样本 $X_1, \cdots, X_m$，相互独立同分布，分布函数为 $F(x)$ 及 $Y_1, \cdots, Y_n$ 相互独立同分布，分布函数为 $F(x-\Delta)$. 现检验假设检验问题

$$H_0: \Delta = 0, \qquad H_1: \Delta > 0.$$

可用 Wilcoxon 统计量作检验.

定理 3.4 在 H_0: $F(x) \equiv G(x)$ 下，Wilcoxon 统计量 W 的分布为

$$P\{W = d\} = \frac{t_{m,n}(d)}{\binom{N}{n}}, \tag{1.3.9}$$

$$d=\frac{n(n+1)}{2},\frac{n(n+1)}{2}+1,\cdots,\frac{n(2m+n+1)}{2},$$

其中 $N=n+m$，$t_{m,n}(d)$ 表示从 $1,2,\cdots,m+n$ 中取 n 个数其和恰为 d 的可能取法的个数.

证明 在 H_0 下，$X_1,\cdots,X_m;Y_1,\cdots,Y_n$ 混合样本的秩向量 $R=(Q_1,\cdots,Q_m,R_1,\cdots,R_n)$ 在集合

$$\mathscr{R}=\{(1,2,\cdots,m+n)\text{ 的全部排列}\}$$

上均匀分布. 故由古典概率公式即得定理结论. 证毕.

计算 $t_{m,n}(d)$，可利用下列递推公式

$$t_{m,n}(d)=t_{m,(n-1)}(d-m-n)+t_{(m-1),n}(d),$$

及初始条件：

$t_{i,0}(0)=1$；

$t_{i,0}(d)=0$， 当 $d\neq 0$，$i=1,\cdots,m$；

$t_{0,j}\left(\frac{j(j+1)}{2}\right)=1$；

$t_{0,j}(d)=0$， 当 $d\neq\frac{j(j+1)}{2}$，$j=1,\cdots,n$.

在常用统计表中，对较小的 m,n 有 W 的检验用临界值表. m，n 较大时可用正态近似(参见第二章).

定理 3.5 在 H_0：$F(x)\equiv G(x)$ 下，Wilcoxon 统计量 W 有

$$E(W)=\frac{n(N+1)}{2},\qquad\qquad (1.3.10)$$

$$\operatorname{var}(W)=mn(N+1)/12,$$

且 W 的分布关于 $n(N+1)/2$ 对称，其中 $N=m+n$.

证明 $E(W)=\sum\limits_{i=1}^{n}E(R_i)=n\cdot\frac{N+1}{2}$，

$$\begin{aligned}\operatorname{var}(W)&=\sum_{i=1}^{n}\operatorname{var}(R_i)+\sum_{i\neq j}\sum\operatorname{cov}(R_i,R_j)\\&=n\cdot\frac{(N+1)(N-1)}{12}-n(n-1)\frac{N+1}{12}\end{aligned}$$

$$= \frac{mn(N+1)}{12}.$$

下面证明 W 的分布关于 $n(N+1)/2$ 对称. 考虑

$$-X_1, \cdots, -X_m, -Y_1, \cdots, -Y_n,$$

这个混合样本在 H_0 下也是独立同分布的,其秩向量在

$$\mathscr{R} = \{(1,2,\cdots,N) \text{ 的全部排列}\}$$

上有均匀分布,这个秩向量为

$$(N+1-Q_1, \cdots, N+1-Q_m, N+1-R_1, \cdots, N+1-R_n),$$

所以

$$\begin{aligned}
&(Q_1, \cdots, Q_m, R_1, \cdots, R_n) \\
&\quad \overset{d}{=\!=} (N+1-Q_1, \cdots, N+1-Q_m, \\
&\qquad N+1-R_1, \cdots, N+1-R_n),
\end{aligned}$$

故

$$W = \sum_{i=1}^{n} R_i \overset{d}{=\!=} \sum_{i=1}^{n} (N+1-R_i) = n(N+1) - W,$$

即有

$$W - \frac{n(N+1)}{2} \overset{d}{=\!=} \frac{n(N+1)}{2} - W.$$

所以 W 的分布关于 $n(N+1)/2$ 对称. 证毕.

Wilcoxon 秩和统计量与 Mann-Whitney 统计量

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \psi(Y_j - X_i) \tag{1.3.11}$$

有下列关系

$$W = U + \frac{n(n+1)}{2}. \tag{1.3.12}$$

统计量 U 表示 (X_i, Y_j) $(i=1,\cdots,m, j=1,\cdots,n)$ 共 mn 个数对中, Y 比 X 大的个数. 上面的关系式可如下证明. 记样本 $Y_1, \cdots, Y_n$ 的次序统计量为 $Y_{(1)} \leqslant \cdots \leqslant Y_{(n)}$, 对应的 $R_1, \cdots, R_n$ 的次序统计量记为 $R_{(1)} \leqslant \cdots \leqslant R_{(n)}$, 则有

$$\#\{X_i < Y_{(1)},\ i = 1,\cdots,m\} = R_{(1)} - 1,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\#\{X_i < Y_{(j)},\ i = 1,\cdots,m\} = R_{(j)} - j,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\#\{X_i < Y_{(n)},\ i = 1,\cdots,m\} = R_{(n)} - n,$$

其中$\#\{A\}$表示集合 A 中元素的个数. 所以

$$U = \sum_{j=1}^{n}(R_{(j)} - j) = \sum_{j=1}^{n}R_j - \sum_{j=1}^{n}j = W - \frac{n(n+1)}{2}.$$

例 2 设有样本 $X_1,\cdots,X_m$ 相互独立同分布，分布函数为 $F(x)$，及样本 $Y_1,\cdots,Y_n$ 相互独立同分布，分布函数为 $F(x-\Delta)$，利用 Wilcoxon 秩和统计量构造 Δ 的区间估计. 由给定的条件，知

$$X_1,\cdots,X_m,Y_1-\Delta,\cdots,Y_n-\Delta$$

独立同分布，其对应的秩向量记为

$$(Q_1,\cdots,Q_m,R_1,\cdots,R_n).$$

由统计量 W 的临界值表可找到 ω_1 和 ω_2 使

$$\begin{aligned}1-\alpha &= P\left\{\omega_1 \leqslant \sum_{i=1}^{n}R_i \leqslant \omega_2\right\}\\ &= P\left\{\omega_1 - \frac{n(n+1)}{2} \leqslant \sum_{i=1}^{m}\sum_{j=1}^{n}\psi(Y_j - \Delta - X_i)\right.\\ &\qquad\qquad \left.\leqslant \omega_2 - \frac{n(n+1)}{2}\right\}.\end{aligned}$$

记 $\omega_1 - \frac{n(n+1)}{2} = K_1$，$\omega_2 - \frac{n(n+1)}{2} = K_2$，$Y_j - X_i$ $(i=1,\cdots,m, j=1,\cdots,n)$共 mn 个变量的次序统计量为 $Z_{(1)} \leqslant \cdots \leqslant Z_{(mn)}$，则由

$$K_1 \leqslant \sum_{i=1}^{m}\sum_{j=1}^{n}\psi(Y_j - X_i - \Delta) \leqslant K_2$$

可解得等价的关系式

$$Z_{(mn-K_1+1)} - \Delta > 0 \text{ 且 } Z_{(mn-K_2)} - \Delta \leqslant 0.$$

故 Δ 的置信度是 $1-\alpha$ 的区间估计为

$$(Z_{(mn-K_2)},\ Z_{(mn-K_1+1)}).$$

§1.4 符号秩统计量

定义 4.1 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布 $F(x)$ 连续，关于 0 点对称. 计数统计量 $\Psi_i=\psi(X_i)$. 随机变量 $|X_1|,\cdots,|X_n|$ 对应的秩向量记为 $(R_1^+,\cdots,R_n^+)$，R_i^+ 称为 X_i 的**绝对秩**，$\Psi_i R_i^+$ 称为 X_i 的**符号秩**.

定理 4.1 若 X 是连续随机变量，分布关于 0 点对称，则随机变量 $|X|$ 与计数统计量 $\psi(X)$ 独立.

证明 对 $t\leqslant 0, i=0$ 或 1，显然有

$$P\{\psi(X)=i,\ |X|\leqslant t\}=P\{\psi(X)=i\}\cdot P\{|X|\leqslant t\}=0.$$

对 $t>0$，有

$$\begin{aligned}P\{\psi(X)=1,|X|\leqslant t\}&=P\{X>0,|X|\leqslant t\}\\&=P\{0<X\leqslant t\}\\&=\frac{1}{2}[P\{0<X\leqslant t\}+P\{-t\leqslant X<0\}]\\&=\frac{1}{2}P\{|X|\leqslant t\}\\&=P\{\psi(X)=1\}\cdot P\{|X|\leqslant t\}.\end{aligned}$$

类似地可证明

$$P\{\psi(X)=0,|X|\leqslant t\}=P\{\psi(X)=0\}\cdot P\{|X|\leqslant t\}.$$

所以 $\psi(X)$ 与 $|X|$ 独立. 证毕.

定理 4.2 若随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布连续，关于 0 点对称，其绝对秩向量 $R^+=(R_1^+,\cdots,R_n^+)$，计数统计量 $\Psi_i=\psi(X_i)$ $(i=1,\cdots,n)$，则 $\Psi_1,\cdots,\Psi_n,R^+$ 相互独立；Ψ_i 有参数为 1/2 的两点分布；R^+ 在集合 $\mathscr{R}=\{(1,\cdots,n)$的排列$\}$上均匀分布.

证明 由定理 4.1、定理 2.1 和定理 3.1 立即可得定理结论. 证毕.

由定理 4.2 我们可知：只依赖于 $\Psi_1,\cdots,\Psi_n,R^+$ 的统计量

$S(\Psi_1,\cdots,\Psi_n,R^+)$，对于关于 0 点对称的连续分布构成的分布类是适应任意分布的.

定理 4.3 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布连续，关于 0 点对称. 令计数统计量 $\Psi_i=\psi(X_i)$ $(i=1,\cdots,n)$，

$$Q=\sum_{i=1}^{n}\Psi_i \quad (\text{表示 } X_1,\cdots,X_n \text{ 中取正值的个数}).$$

又当 $Q=q$ 时，$X_1,\cdots,X_n$ 中取正值者对应的绝对秩记为 $S_1<\cdots<S_q$，则

$$P\{Q=q,S_1=t_1,\cdots,S_q=t_q\}=\begin{cases}\left(\dfrac{1}{2}\right)^n, & \text{当 } q=0,1,\cdots,n,\text{且 } t_1\leqslant\cdots\leqslant t_q,\\ 0, & \text{其他}.\end{cases} \tag{1.4.1}$$

证明 (1.4.1)式中的其他情况结果为 0 是显然的.

当 $q=0,1,\cdots,n$，且 $1\leqslant t_1\leqslant\cdots\leqslant t_q\leqslant n$ 时，

$$\begin{aligned}
&P\{Q=q,S_1=t_1,\cdots,S_q=t_q\}\\
&=\sum^{\binom{n}{q}}P\{\Psi_{d_1}=1,\cdots,\Psi_{d_q}=1,\Psi_{d_{(q+1)}}=0,\cdots,\Psi_{d_n}=0,\\
&\qquad S_1=t_1,\cdots,S_q=t_q\}\\
&=\sum^{\binom{n}{q}}P\{S_1=t_1,\cdots,S_q=t_q\mid\Psi_{d_1}=1,\cdots,\Psi_{d_q}=1,\\
&\qquad \Psi_{d_{(q+1)}}=0,\cdots,\Psi_{d_n}=0\}\\
&\qquad\cdot P\{\Psi_{d_1}=1,\cdots,\Psi_{d_q}=1,\Psi_{d_{(q+1)}}=0,\cdots,\Psi_{d_n}=0\}\\
&=\sum^{\binom{n}{q}}\sum^{q!(n-q)!}P\{(R^+_{d_1},\cdots,R^+_{d_q})=t,(R^+_{d_{(q+1)}},\cdots,R^+_{d_n})=r\mid\\
&\qquad \Psi_{d_1}=1,\cdots,\Psi_{d_q}=1,\Psi_{d_{(q+1)}}=0,\cdots,\Psi_{d_n}=0\}\\
&\qquad\cdot\left(\frac{1}{2}\right)^n
\end{aligned}$$

$$=\sum^{\binom{n}{q}}\sum^{q!(n-q)!} P\{(R_{d_1}^+,\cdots,R_{d_n}^+)=(t,r)\}\cdot\left(\frac{1}{2}\right)^n$$

$$=\left(\frac{1}{2}\right)^n\sum^{\binom{n}{q}}\sum^{q!(n-q)!}\frac{1}{n!}=\left(\frac{1}{2}\right)^n,$$

其中$\sum^{\binom{n}{q}}$表示对从$(1,\cdots,n)$中取q个$(d_1,\cdots,d_q)$的一切可能求和；t表示$(t_1,\cdots,t_q)$的一个排列，r表示$(1,\cdots,n)$中除去$(t_1,\cdots,t_q)$后，所剩的$n-q$个数的一个排列；$\sum^{q!(n-q)!}$表示对上述两个排列的一切可能求和. 证毕.

可以用符号秩统计量构造关于对称点问题的检验.

设随机变量$X_1,\cdots,X_n$相互独立同分布，分布$F(x-\theta)$连续，$F(t)$关于0点对称. 检验假设

$$H_0:\ \theta=\theta_0,\qquad H_1:\ \theta>\theta_0.$$

令$Z_i=X_i-\theta_0$. $Z_1,\cdots,Z_n$的符号秩记为$\Psi_1R_1^+,\cdots,\Psi_nR_n^+$. Wilcoxon 符号秩检验统计量为

$$W^+=\sum_{i=1}^{n}\Psi_iR_i^+,\tag{1.4.2}$$

其否定域形式为

$$W^+\geqslant\omega^+(\alpha,n).$$

给定检验水平α，由W^+的临界值表可确定临界值$\omega^+(\alpha,n)$.

例 1 设进行n次对比试验，得数据

$$(x_1,y_1),\cdots,(x_n,y_n).$$

x_i是对比试验的结果. y_i是施加处理的试验的结果. 它们分别来自连续分布总体，可认为两者差一个处理效应. 现检验有无处理效应θ，即检验

$$H_0:\ \theta=0,\qquad H_1:\ \theta>0.$$

考虑$y_i-x_i(i=1,\cdots,n)$，在对比试验中它们可认为是一个对称分布的一组样本值. 在H_0下对称分布的对称点为0，故可用

Wilcoxon 符号秩检验上述假设.

定理 4.4 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布,分布连续且关于 0 点对称,其对应的符号秩为 $\Psi_1R_1^+,\cdots,\Psi_nR_n^+$,则统计量 $W^+=\sum_{i=1}^n\Psi_iR_i^+$ 的分布为

$$P\{W^+=k\}=\begin{cases}c_n(k)/2^n, & k=0,1,\cdots,\dfrac{n(n+1)}{2},\\ 0, & \text{其他},\end{cases} \tag{1.4.3}$$

其中 $c_n(k)$是和数恰为 k 的$\{1,\cdots,n\}$的子集的个数.

证明 由定理 4.3,当 $k=0,1,\cdots,\frac{n(n+1)}{2}$时,

$$\begin{aligned}P\{W^+=k\}&=\sum_{q=0}^n\sum_{t_1+\cdots+t_q=k}P\{Q=q,S_1=t_1,\cdots,S_q=t_q\}\\&=c_n(k)/2^n.\end{aligned}$$

其他情形概率为 0 是显然的. 证毕.

对于 $c_n(k)$有下列递推式和初始条件:

$$c_n(k)=c_{n-1}(k-n)+c_{n-1}(k),$$

$$c_1(0)=c_1(1)=1,\quad c_1(d)=0,\quad \text{当 } d\neq 0 \text{ 或 } 1.$$

定理 4.5 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布,分布连续且关于 0 点对称,其对应的符号秩为 $\Psi_1R_1^+,\cdots,\Psi_nR_n^+$,统计量 $W^+=\sum_{i=1}^n\Psi_iR_i^+$,则

$$E(W^+)=\frac{n(n+1)}{4}, \tag{1.4.4}$$

$$\operatorname{var}(W^+)=\frac{n(n+1)(2n+1)}{24}, \tag{1.4.5}$$

且 W^+的分布关于$\frac{n(n+1)}{4}$对称.

证明 由定理 4.2,可得

$$E(\Psi_i R_i^+)=E(\Psi_i)\cdot E(R_i^+)=\frac{1}{2}\cdot\frac{n+1}{2}=\frac{n+1}{4}\quad(i=1,\cdots,n),$$

$$\begin{aligned}\mathrm{var}(\Psi_i R_i^+) &= E(\Psi_i R_i^+ - E(\Psi_i)\cdot E(R_i^+))^2 \\ &= E(\Psi_i^2)\cdot E(R_i^+ - E(R_i^+))^2 + (E(R_i^+))^2 \\ &\quad \cdot E(\Psi_i - E(\Psi_i))^2 \\ &= \frac{1}{2}\cdot\frac{(n+1)(n-1)}{12} + \left(\frac{n+1}{2}\right)^2\cdot\frac{1}{4} \\ &= \frac{(n+1)(5n+1)}{48},\quad i=1,\cdots,n.\end{aligned}$$

$$\begin{aligned}\mathrm{cov}(\Psi_i R_i^+,\Psi_j R_j^+) &= E(\Psi_i)\cdot E(\Psi_j)\cdot\mathrm{cov}(R_i^+,R_j^+) \\ &= -\frac{(n+1)}{48},\quad i\neq j.\end{aligned}$$

故有

$$\begin{aligned}\mathrm{var}(W^+) &= \sum_{i=1}^{n}\mathrm{var}(\Psi_i R_i^+) + \sum_{i\neq j}\sum\mathrm{cov}(\Psi_i R_i^+,\Psi_j R_j^+) \\ &= n\cdot\frac{(n+1)(5n+1)}{48} - \frac{n(n-1)(n+1)}{48} \\ &= \frac{n(n+1)(2n+1)}{24}.\end{aligned}$$

又当分布关于 0 点对称时，有

$$\Psi_i \xlongequal{d} 1-\Psi_i\quad(i=1,\cdots,n),$$

从而

$$W^+=\sum_{i=1}^{n}\Psi_i R_i^+ \xlongequal{d} \sum_{i=1}^{n}(1-\Psi_i)R_i^+ = \frac{n(n+1)}{2}-W^+.$$

所以 W^+ 的分布关于 $\frac{n(n+1)}{4}$ 对称. 证毕.

W^+ 统计量也有一个计数统计量形式的表达式. 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，计数统计量 $\Psi_i=\phi(X_i)$ $(i=1,\cdots,n)$，$(X_1,\cdots,X_n)$ 的秩向量为 $(R_1,\cdots,R_n)$，而绝对秩向量为 $(R_1^+,\cdots,R_n^+)$，则有关系式

$$\Psi_i R_i^+ = \sum_{j \leqslant R_i} \psi(X_{(R_i)} + X_{(j)}), \quad i = 1, \cdots, n.$$

从而

$$\begin{aligned} W^+ &= \sum_{i=1}^{n} \Psi_i R_i^+ = \sum_{i=1}^{n} \sum_{j \leqslant R_i} \psi(X_{(R_i)} + X_{(j)}) \\ &= \sum_{i=1}^{n} \sum_{j \leqslant i} \psi(X_{(i)} + X_{(j)}) = \sum_{i=1}^{n} \sum_{j \leqslant i} \psi(X_i + X_j). \end{aligned} \tag{1.4.6}$$

例 2 由 W^+ 统计量构造分布对称点的区间估计. 设随机变量 $X_1, \cdots, X_n$ 相互独立同分布，分布 $F(x-\theta)$ 连续，参数 θ 未知，$F(t)$ 关于 0 点对称. 令 $Z_i = X_i - \theta$ $(i = 1, \cdots, n)$，则 Z_i 的分布关于 0 点对称.

$$W^+ = \sum_{i=1}^{n} \sum_{j \leqslant i} \psi(Z_i + Z_j) = \sum_{i=1}^{n} \sum_{j \leqslant i} \psi(X_i + X_j - 2\theta)$$

为 Wilcoxon 符号秩统计量. 查其临界值表可得 ω_1, ω_2 使得

$$P\{\omega_1 \leqslant W^+ \leqslant \omega_2\} = 1 - \alpha.$$

将 $\left\{\dfrac{X_i + X_j}{2}, j \leqslant i, i, j = 1, \cdots, n\right\}$ 共 $\dfrac{n(n+1)}{2}$ 个变量的次序统计量记为 $V_{(1)} \leqslant V_{(2)} \leqslant \cdots \leqslant V_{(n(n+1)/2)}$，则有

$$\begin{aligned} \{\omega_1 \leqslant W^+ \leqslant \omega_2\} &= \left\{\omega_1 \leqslant \sum_{i=1}^{n} \sum_{j \leqslant i} \psi\left(\frac{X_i + X_j}{2} - \theta\right) \leqslant \omega_2\right\} \\ &= \{V_{\left(\frac{n(n+1)}{2} - \omega_1 + 1\right)} - \theta > 0 \text{ 且 } V_{\left(\frac{n(n+1)}{2} - \omega_2\right)} - \theta \leqslant 0\}. \end{aligned}$$

故 θ 的置信度为 $1-\alpha$ 的区间估计为

$$(V_{\left(\frac{n(n+1)}{2} - \omega_2\right)}, V_{\left(\frac{n(n+1)}{2} - \omega_1 + 1\right)}).$$

§ 1.5 拟秩统计量

1963 年 L. Moses 提出一种利用秩的方法. 其秩不是原始数据的秩，而是原始数据的某种函数的秩. 当然函数要选择得便于应

用. 例如, $g(x_1,\cdots,x_N)$是$(x_1,\cdots,x_N)$的对称函数(即对$(1,\cdots,N)$的任意排列$(\alpha_1,\cdots,\alpha_N)$, 有 $g(x_1,\cdots,x_N)=g(x_{\alpha_1},\cdots,x_{\alpha_N})$). 令

$$V_i = h(X_i,g(X_1,\cdots,X_N)), \quad i=1,\cdots,N,$$

其中 $h(x,g)$是一给定的函数. 则当随机变量 $X_1,\cdots,X_N$ 相互独立同分布时, $V=(V_1,\cdots,V_N)$是可交换的随机向量, 即对任一个$(1,\cdots,N)$的排列$(\alpha_1,\cdots,\alpha_N)$, 有

$$(V_{\alpha_1},\cdots,V_{\alpha_N}) \xlongequal{d} (V_1,\cdots,V_N).$$

我们注意到本章定义 3.1 及定理 3.1 中, $Z_1,\cdots,Z_N$ 独立同分布的条件可减弱为$(Z_1,\cdots,Z_N)$是可交换的随机向量, 故当 V_i 的分布连续时, 则由 $V_1,\cdots,V_N$ 确定的秩向量 $R(V)=(R_1(V),\cdots,R_N(V))$ 也在集合

$$\mathscr{R} = \{(1,\cdots,N)\text{ 的全部排列}\}$$

上均匀分布, 因而可利用 $R(V)$构成非参数适应任意分布的统计量.

例 1 设随机变量 $X_1,\cdots,X_m$ 相互独立同分布, 分布函数为 $F(x-\theta)$; 随机变量 $Y_1,\cdots,Y_n$ 相互独立同分布, 分布为 $F\left(\frac{x-\theta}{\eta}\right)$, 分布函数 $F(x)$连续, $F(0)=\frac{1}{2}$, $-\infty<\theta<\infty$, $\eta>0$. 对假设检验问题

$$H_0:\ \eta=1, \qquad H_1:\ \eta>1.$$

当 θ 未知时, 可如下构造拟秩统计量检验. 令

$$\hat{\theta} = \text{median}(X_1,\cdots,X_m,Y_1,\cdots,Y_n), \tag{1.5.1}$$

$$V_i = \begin{cases} |X_i-\hat{\theta}|, & i=1,\cdots,m, \\ |Y_{(i-m)}-\hat{\theta}|, & i=m+1,\cdots,m+n, \end{cases} \tag{1.5.2}$$

当 $N=m+n$ 为奇数时, V_i 的分布连续. 用 $V_1,\cdots,V_N$ 的秩向量 $R(V)=(R_1(V),\cdots,R_N(V))$构造 Wilcoxon 秩和检验, 检验统计量为

$$W = \sum_{i=m+1}^{N} R_i(V),$$

可用来检验上述假设. 当 N 为偶数时,有两个 V_i 相等,形成所谓"结",这涉及结的处理技术. 可利用$(V_1,\cdots,V_N)$在给定$(V_1,\cdots,V_N)=(v_1,\cdots,v_N)$的条件下,$R(V)$在

$$\mathscr{R}'=\left\{\left(1,\cdots,\frac{N}{2}-1,\frac{N}{2}+\frac{1}{2},\frac{N}{2}+\frac{1}{2},\frac{N}{2}+2,\cdots,N\right)\text{的全部排列}\right\}$$

上均匀分布. 由此给定显著水平可确定一条件否定域.

例 2 两条回归线的比较. 设

$$(X_{11},Y_{11}),\cdots,(X_{1m},Y_{1m});$$
$$(X_{21},Y_{21}),\cdots,(X_{2n},Y_{2n})$$

为两组独立样本. 满足 X_{ij}给定后,Y_{ij}的分布为

$$P\{Y_{ij}\leqslant y|X_{ij}=x_{ij}\}=G(y-\alpha-\beta_i x_{ij})$$
$$i=1,2,\quad j=1,\cdots,m\text{ 或 }n,$$

其中 $G(t)$是一个连续分布函数. 现检验

$$H_0:\beta_1=\beta_2(=\beta),\qquad H_1:\beta_2>\beta_1.$$

记

$$(U_k,V_k)=\begin{cases}(X_{1k},Y_{1k}), & k=1,\cdots,m,\\ (X_{2(k-m)},Y_{2(k-m)}), & k=m+1,\cdots,m+n.\end{cases}$$

这些数据可看作一对随机变量(U,V)的样本值,在 H_0 下,由他们确定的 V 对 U 的线性回归系数为

$$\hat{\alpha}=\overline{V}-\beta\overline{U},\quad \overline{V}=\frac{1}{N}\sum_{k=1}^{N}V_k,\quad \overline{U}=\frac{1}{N}\sum_{k=1}^{N}U_k,\tag{1.5.3}$$

$$\hat{\beta}=\sum_{k=1}^{N}(U_k-\overline{U})(V_k-\overline{V})\Big/\sum_{k=1}^{N}(U_k-\overline{U})^2,\quad N=m+n.\tag{1.5.4}$$

$\hat{\alpha},\hat{\beta}$ 均为(U_k,V_k) $(k=1,\cdots,N)$的对称函数. 今令

$$W_k=[V_k-\hat{\alpha}-\hat{\beta}U_k]\cdot\operatorname{sgn}(U_k),\quad k=1,\cdots,N,\tag{1.5.5}$$

其中

$$\operatorname{sgn}(u)=\begin{cases}1, & \text{当 } u>0 \text{ 时}, \\ 0, & \text{当 } u=0 \text{ 时}, \\ -1, & \text{当 } u<0 \text{ 时}.\end{cases}$$

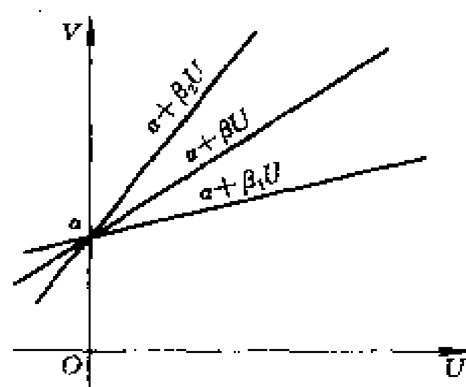


图 1

由于给定 $X_{ij}=x_{ij}$, $Y_{1j}-\alpha-\beta_1x_{1j}$ $(j=1,\cdots,m)$ 以及 $Y_{2j}-\alpha-\beta_2x_{2j}$ $(j=1,\cdots,n)$ 具有相同分布 $G(y)$. 而两组样本混合的回归线与各自分别的回归线三者有图 1 中的关系(当 $\beta_2>\beta_1$ 时),即应有 $\beta_1<\beta<\beta_2$,从而当 $x_{ij}>0$ 时,

$$Y_{2j}-\alpha-\beta x_{2j}=Y_{2j}-\alpha-\beta_2x_{2j}+(\beta_2-\beta)x_{2j},\quad j=1,\cdots,n$$

有分布 $G(y-(\beta_2-\beta)x_{2j})$,而

$$Y_{1j}-\alpha-\beta x_{1j}=Y_{1j}-\alpha-\beta_1x_{1j}-(\beta-\beta_1)x_{1j},\quad j=1,\cdots,m$$

有分布 $G(y+(\beta-\beta_1)x_{ij})$. $Y_{2j}-\alpha-\beta x_{2j}$随机地大于 $Y_{1j}-\alpha-\beta x_{1j}$. 当 $x_{ij}<0$ 时,$-(Y_{2j}-\alpha-\beta x_{2j})$随机地大于$-(Y_{1j}-\alpha-\beta x_{1j})$,故取 $W_1,\cdots,W_N$ 的秩构成 Wilcoxon 秩和统计量

$$T=\sum_{j=m+1}^{N}R_j(W),$$

则 T 值大有利备择假设 H_1: $\beta_2>\beta_1$,其中$(R_1(W),\cdots,R_m(W),R_{m+1}(W),\cdots,R_N(W))$为$(W_1,\cdots,W_m,W_{m+1},\cdots,W_N)$对应的秩向量. 利用 Wilcoxon 秩和分布对给定的显著水平可确定出否定域的临界值.

§1.6 条件的适应任意分布的检验

一般一个检验问题当总体分布 $F(x)$不同时,"好的"统计量也不同,虽然总体分布 $F(x)$是未知的,但从观测的样本数据可以对总体有一点了解,因此设想在进行检验前,依据所得的样本观测数据对模型作一些判断,然后根据判断结果挑选适当的检验统计量.

当然按照假设检验的传统观点，要保证对所有给定判断条件的情况均有显著水平 α. 下列定理保证这样的检验总的显著水平仍然是 α.

定理 6.1 设 $\mathscr{F}$ 是一个分布类，对关于 $\mathscr{F}$ 的假设 H_0，有 k 个适应任意分布的检验，对应的统计量为 $S_1,\cdots,S_k$，相应的否定域为 $C_1,\cdots,C_k$，即有

$$P\{S_j \in C_j | H_0, F(x) \in \mathscr{F}\} = \alpha, \quad j = 1,\cdots,k. \tag{1.6.1}$$

若存在统计量 V，当 $F(x)\in\mathscr{F}$，且 H_0 成立时，V 与 $(S_1,\cdots,S_k)$ 独立，V 之值域可分成两两分离的集 $D_1,\cdots,D_k$，则下述关于 H_0 的检验也是对 $\mathscr{F}$ 为适应任意分布的，显著性水平为 α 的检验：

若 $V\in D_j$ $(j=1,\cdots,k)$，则当 $S_j\in C_j$ 时，拒绝原假设 H_0，接受备择假设 H_1.

证明

$$\begin{aligned}
&P\{\text{拒绝 } H_0 | H_0, F(x) \in \mathscr{F}\} \\
&= P\left\{ \bigcup_{j=1}^{k} \{V \in D_j, S_j \in C_j\} | H_0, F(x) \in \mathscr{F} \right\} \\
&= \sum_{j=1}^{k} P\{V \in D_j, S_j \in C_j | H_0, F(x) \in \mathscr{F}\} \\
&= \sum_{j=1}^{k} P\{V \in D_j | H_0, F(x) \in \mathscr{F}\} \\
&\qquad \cdot P\{S_j \in C_j | H_0, F(x) \in \mathscr{F}\} \\
&= \alpha \sum_{j=1}^{k} P\{V \in D_j | H_0, F(x) \in \mathscr{F}\} \\
&= \alpha.
\end{aligned}$$

证毕.

当用秩统计量作为检验统计量时，利用下面的定理 6.2 很容易构造出条件的适应任意分布的否定域.

定理 6.2 设随机变量 $X_1,\cdots,X_N$ 相互独立同分布，其次序

统计量为 $X_{(1)}\leqslant\cdots\leqslant X_{(N)}$，对应的秩向量为

$$R=(R_1,\cdots,R_N),$$

则 $R=(R_1,\cdots,R_N)$ 与 $X^0=(X_{(1)},\cdots,X_{(N)})$ 相互独立.

证明 对 $(1,\cdots,N)$ 的任意一个排列

$$r=(r_1,\cdots,r_N)$$

和 N 维实数集合

$$A\subset\{(x_1,\cdots,x_N)\,|\,x_1\leqslant\cdots\leqslant x_N\},$$

$$\begin{aligned}P\{R=r,X^0\in A\}&=P\{X_1=X_{(r_1)},\cdots,X_N=X_{(r_N)},X^0\in A\}\\&=P\{(X_{d_1},\cdots,X_{d_N})\in A\}\\&=P\{(X_1,\cdots,X_N)\in A\},\end{aligned}$$

其中 $d_j=i$ 当 $r_i=j$，对一切 i,j. 而

$$\begin{aligned}P\{X^0\in A\}&=\sum_{(r_1,\cdots,r_N)}P\{R=r,X^0\in A\}\\&=N!P\{(X_1,\cdots,X_N)\in A\},\end{aligned}$$

故

$$P\{R=r,X^0\in A\}=\frac{1}{N!}P\{X^0\in A\}=P\{R=r\}P\{X^0\in A\}.$$

证毕.

由本定理知，给定 $(X_{(1)},\cdots,X_{(N)})=(x_{(1)},\cdots,x_{(N)})$，则 $(X_1,\cdots,X_N)$ 在 $(x_{(1)},\cdots,x_{(N)})$ 的 $N!$ 个排列点上均匀分布，即

$$\begin{aligned}\frac{1}{N!}&=P\{R=r\,|\,X^0=(x_{(1)},\cdots,x_{(N)})\}\\&=P\{X_1=X_{(r_1)},\cdots,X_N=X_{(r_N)}\,|\,X^0=(x_{(1)},\cdots,x_{(N)})\}.\end{aligned}$$

由此可构造条件的适应任意分布的检验.

例 1 设 $x_1=4.3,x_2=6.0,x_3=3.6$ 是来自总体 $F(x)$ 的样本观测值；$y_1=7.4,y_2=5.5,y_3=6.2$ 是来自总体 $F(x-\Delta)$ 的样本观测值. 它们的组合在一起的混合排序为

$$z_{(1)}=x_3=3.6,\ z_{(2)}=x_1=4.3,\ z_{(3)}=y_2=5.5,$$

$$z_{(4)}=x_2=6.0,\ z_{(5)}=y_3=6.2,\ z_{(6)}=y_1=7.4.$$

为检验

$$H_0: \Delta = 0, \qquad H_1: \Delta > 0.$$

考虑统计量

$$S(X_1, X_2, X_3, Y_1, Y_2, Y_3) = \overline{Y} - \overline{X}. \tag{1.6.2}$$

给定$(z_{(1)}, z_{(2)}, z_{(3)}, z_{(4)}, z_{(5)}, z_{(6)})$，$\overline{Y}-\overline{X}$的取值共有$C_6^3=20$个，记为$s_1 \leqslant s_2 \leqslant \cdots \leqslant s_{20}$，则

$$P\{\overline{Y} - \overline{X} = s_j \mid \text{给定}(z_{(1)}, z_{(2)}, \cdots, z_{(6)})\} = \frac{1}{20}, \quad j = 1, \cdots, 20.$$

给定显著性水平α，用上述分布可确定否定域. 对于假设检验问题H_0: $\Delta=0$, H_1: $\Delta>0$，$\overline{Y}-\overline{X}$大时应否定$H_0$. 若取$\alpha=0.10$，则可定出否定域为

$$\{s_{20} = 2.067,\ s_{19} = 1.733\}.$$

例 2 对两样本问题，当总体分布$F(x)$为对称分布、左偏斜分布、右偏斜分布，或分布在远离中心位置的概率较大(所谓重尾分布)，或分布在远离中心位置的概率较小(所谓轻尾分布)等各种不同类型分布时，有各种不同的适宜的秩统计量的检验(参见第三章)，它们都是适应任意分布的非参数检验. 当我们能对总体的分布类型有所了解时，可选用适宜的检验统计量，从而可提高功效.

当两样本来自相同的总体时(即两样本问题的H_0成立)，从定理 6.2 知秩统计量与次序统计量独立，因而利用次序统计量作前一步的判断，不会影响后一步用秩统计量作检验的显著性水平. 因此，有人用

$$Q_3 = \frac{\overline{U}_{0.05} - \overline{M}_{0.5}}{\overline{M}_{0.5} - \overline{L}_{0.05}} \tag{1.6.3}$$

来度量偏倚，取

$$Q_4 = \frac{\overline{U}_{0.05} - \overline{L}_{0.05}}{\overline{U}_{0.5} - \overline{L}_{0.5}} \tag{1.6.4}$$

来度量轻尾重尾，其中$\overline{U}_\alpha$, $\overline{M}_\alpha$, $\overline{L}_\alpha$分别表示两样本的组合样本次序统计量中最大，中间，最小的αN个次序统计量观测值的平均值，

$N=m+n$ 为两样本的样本容量之和.

例 3 条件否定域的一个重要应用是处理样本观测的结(tie),即处理有观测值相等的情况. 例如设 $x_1=10.4, x_2=7.8, x_3=8.4$ 来自总体 $F(x)$; $y_1=12.6, y_2=8.4$ 来自总体 $F(x-\Delta)$,检验

$$H_0: \Delta = 0, \qquad H_1: \Delta > 0.$$

将 x_1, x_2, x_3, y_1, y_2 组合在一起排序,则 x_3, y_2 都等于 8.4 形成一个结,在确定它们的秩时,只好取它们的平均秩,即 $(x_1, x_2, x_3, y_1, y_2)$ 对应的秩向量记为 $(Q_1, Q_2, Q_3, R_1, R_2)$ 时,

$x_2=7.8$ 的秩为 $Q_2=1$;

$x_3=8.4$ 的秩为 $Q_3=\dfrac{2+3}{2}=2.5$;

$y_2=8.4$ 的秩为 $R_2=\dfrac{2+3}{2}=2.5$;

$x_1=10.4$ 的秩为 $Q_1=4$;

$y_1=12.6$ 的秩为 $R_1=5$,

则 Wilcoxon 秩和统计量的观测值

$$W = R_1 + R_2 = 5 + 2.5 = 7.5.$$

此时,用 § 1.3 节中 H_0 下的结论: $R=(Q_1, Q_2, Q_3, R_1, R_2)$ 均匀分布,显然是不行的. 利用条件概率,H_0 成立的前提下,在混合样本次序统计量 $(z_{(1)}, z_{(2)}, z_{(3)}, z_{(4)}, z_{(5)})=(7.8, 8.4, 8.4, 10.4, 12.6)$ 的条件下, $(X_1, X_2, X_3, Y_1, Y_2)$ 的秩统计量 R 有分布: 在 $(1, 2.5, 2.5, 4, 5)$ 的任一排列上有概率 $1/5!$. 从而确定出 W 的分布概率为

W	3.5	5	6.5	7.5	9
P	$\dfrac{2}{10}$	$\dfrac{3}{10}$	$\dfrac{2}{10}$	$\dfrac{2}{10}$	$\dfrac{1}{10}$

由此可确定出一个条件否定域.

例 4 条件适应任意分布的非参数检验的一个著名情况是 Fisher 的 2×2 列联表精确检验.

设有Ⅰ、Ⅱ两个总体，每一总体包含 E_1、E_2 两类个体，从两总体中共观测 N 个样本单元，记录：

a=来自Ⅰ的 E_1 的个数；

b=来自Ⅰ的 E_2 的个数；

c=来自Ⅱ的 E_1 的个数；

d=来自Ⅱ的 E_2 的个数.

构成列联表：

	E_1	E_2	
Ⅰ	a	b	$a+b$
Ⅱ	c	d	$c+d$
	$a+c$	$b+d$	N

现检验两个总体是否相同，即检验

H_0：Ⅰ，Ⅱ两总体 E_1,E_2 的比例相同；

H_1：Ⅰ中 E_1 的比例大于Ⅱ中 E_1 的比例.

Fisher 使用统计量

$$S = ad - bc,$$

S 大时否定 H_0. 在固定列联表边缘，即给定 $a+b$,$c+d$,$a+c$,$b+d$ 的条件下，可构造条件否定域. 在此条件下 a,b,c,d 四个数中有一个确定，则其余的均被确定. 而在 H_0 下，

$$P\{a=t \mid \text{给定 } a+b, c+d, a+c, b+d\}$$
$$= \frac{\binom{a+c}{t}\binom{b+d}{a+b-t}}{\binom{N}{a+b}}. \tag{1.6.5}$$

从而可确定 $S=ad-bc$ 的分布，按 S 值的大小，从最大值开始，依次排出，当给定显著性水平 α，则可将前面的概率逐步累加，直到适宜的满足 α 水平的某一临界值，此前排出的值将构成条件否定域. 若 S 落入否定域，则否定 H_0.

习　题

1. 考虑回归模型 $Y_i=\beta d_i+\varepsilon_i$ $(i=1,\cdots,n)$，其中 Y_i 是观测的随机变量，β 为未知参数，$0\leqslant d_1\leqslant d_2\leqslant\cdots\leqslant d_n$ 为已知常数，$\varepsilon_1,\cdots,\varepsilon_n$ 为连续随机变量，相互独立同分布，中位数为 0，试对假设检验问题

$$H_0:\ \beta=0,\qquad H_1:\ \beta>0$$

构造 d-free（适应任意分布）的计数检验.

2. 设随机变量 $Z_1,Z_2,\cdots,Z_N$ 相互独立同分布，分布连续，其对应的秩向量为 $R=(R_1,R_2,\cdots,R_N)$，假定 $N\geqslant 2$，令 $V=R_1-R_N$，试证明

$$P\{V=k\}=\begin{cases}\dfrac{N-|k|}{N(N-1)}, & \text{当 } |k|=1,\cdots,N-1 \text{ 时},\\ 0, & \text{其他}.\end{cases}$$

3. 设 $(X_1,\cdots,X_m)$ 和 $(Y_1,\cdots,Y_n)$ 分别是来自连续分布 $F(x)$ 和 $G(x)$ 的样本，混合样本 $X_1,\cdots,X_m,Y_1,\cdots,Y_n$ 的对应的秩向量为

$$R=(Q_1,\cdots,Q_m,R_1,\cdots,R_n).$$

令

$$R_{\max}=\max\{R_1,\cdots,R_n\},$$

说明当 $F(x)\equiv G(x)$ 时，$R_{\max}$ 是 d-free 的.

4. 设随机变量 $X_1,\cdots,X_n$ 是来自关于 0 点对称的连续分布的独立随机样本，W^+ 为其 Wilcoxon 符号秩统计量，证明

$$W^+\overset{d}{=\!=}\sum_{j=1}^{n}V_j,$$

其中 $V_1,\cdots,V_n$ 是相互独立的随机变量，且有

$$P\{V_j=j\}=P\{V_j=0\}=\frac{1}{2},\quad j=1,\cdots,n.$$

5. 在上题的条件下，求 W^+ 的矩母函数，并用它来计算

$E(W^+)$和 $\mathrm{var}(W^+)$.

6. 设 X 是一个随机变量，称 X 关于 c 权对称，当且仅当存在 $\lambda>0$，使对一切 $t>0$，有

$$P\{X > c + t\} = \lambda P\{X < c - t\}.$$

若 X 关于 c 权对称，证明当 $0<P\{X<c\}=P\{X\leqslant c\}<1$ 时，变量 $|X-c|$ 与 $\Psi=\psi(X-c)$ 独立.

7. 设随机变量 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 是分别来自连续分布 $F(x-\theta_X)$和 $F\left(\frac{x-\theta_Y}{\eta}\right)$ 的独立随机样本，其中 $-\infty<\theta_X,\ \theta_Y<+\infty,\eta>0$，且 $F(0)=\frac{1}{2}$. 为了检验假设

$$H_0:\ \eta = 1, \qquad H_1:\ \eta > 1,$$

Moses 提出如下方法：选一正整数 $k\geqslant 2$，随机地将 $X_1,\cdots,X_m$ 分成大小为 k 的 m' 个组，$Y_1,\cdots,Y_n$ 分成大小为 k 的 n' 个组. 以 $X_{i1},\cdots,X_{ik}$记 X 观测的第 i 个组，$Y_{j1},\cdots,Y_{jk}$记 Y 观测的第 j 个组，记

$$C_i = \sum_{l=1}^{k}(X_{il} - \overline{X}_i)^2, \quad \overline{X}_i = \frac{1}{k}\sum_{l=1}^{k}X_{il}, \quad i = 1,\cdots,m',$$

$$D_j = \sum_{l=1}^{k}(Y_{jl} - \overline{Y}_j)^2, \quad \overline{Y}_j = \frac{1}{k}\sum_{l=1}^{k}Y_{jl}, \quad j = 1,\cdots,n',$$

统计量$(C_1,\cdots,C_{m'},D_1,\cdots,D_{n'})$对应的秩向量记为$(Q_1,\cdots,Q_{m'},R_1,\cdots,R_{n'})$，令

$$T = \sum_{j=1}^{n'}R_j,$$

以 T 为统计量检验前述假设，当 $T\geqslant t(\alpha,m',n')$时，拒绝 H_0，$t(\alpha,m',n')$为 T 在 H_0 下的 $1-\alpha$ 分位数. 试回答下列问题：

(a) 论证无论未知的 θ_X 和 θ_Y 为何值，在 H_0 下 T 是 d-free 的.

(b) 在 H_0 下，求 T 的分布? 临界值 $t(\alpha,m',n')$易求吗?

第二章　U 统 计 量

§2.1　一样本 U 统计量

定义 1.1　对分布 F 的参数 θ，如果存在样本量为 r 的样本 $X_1,\cdots,X_r$ 的统计量 $h(X_1,\cdots,X_r)$使

$$E_F h(X_1,\cdots,X_r)=\theta,\quad 对一切 F\in\mathscr{F},$$

则称参数 θ 对分布族 $\mathscr{F}$ 是 **r 可估的**，其中 $E_F h(\cdot)$表示统计量 $h(\cdot)$在总体分布为 F 时的期望值. 使上式成立的最小的 r 称为**可估参数 θ 的自由度**.

这里 $h(X_1,\cdots,X_r)$称为 θ 的**核**(Kernel).

我们不妨认为 $h(x_1,\cdots,x_r)$是对称函数，即对任一$(1,\cdots,r)$的排列$(\alpha_1,\cdots,\alpha_r)$有

$$h(x_{\alpha_1},\cdots,x_{\alpha_r})=h(x_1,\cdots,x_r).$$

这是因为：若 $h(\cdot)$不对称，由于

$$h(X_1,\cdots,X_r)\overset{d}{=\!=}h(X_{\alpha_1},\cdots,X_{\alpha_r}),$$

令

$$h^*(X_1,\cdots,X_r)=\frac{1}{r!}\sum_{(\alpha_1,\cdots,\alpha_r)}h(X_{\alpha_1},\cdots,X_{\alpha_r}),$$

则 $h^*(\cdot)$也满足定义 1.1 的要求，且是对称的.

定义 1.2　设随机变量 $X_1,\cdots,X_n$ 是总体 $F(x)\in\mathscr{F}$ 的样本. r 可估参数 θ 有对称核 $h(X_1,\cdots,X_r)$，则由 $h(\cdot)$形成的统计量

$$U(X_1,\cdots,X_n)=\frac{1}{\binom{n}{r}}\sum_{(\beta_1,\cdots,\beta_r)}h(X_{\beta_1},\cdots,X_{\beta_r})\qquad(2.1.1)$$

称为**参数 θ 的 U 统计量**，其中 $\sum_{(\beta_1,\cdots,\beta_r)}$ 表示对一切$(1,\cdots,n)$中取 r 个的一切组合$(\beta_1,\cdots,\beta_r)$求和.

例 1 设 $\mathscr{F}$ 为全体一阶矩存在的分布类，则均值 $\theta=E_F(X_1)$ 是对 $\mathscr{F}$ 自由度为 1 的可估参数. 对称核为

$$h(X_1)=X_1.$$

由此形成的 U 统计量为

$$U(X_1,\cdots,X_n)=\frac{1}{\binom{n}{1}}\sum_{i=1}^{n}X_i=\overline{X}.$$

例 2 设 $\mathscr{F}$ 为全体二阶矩有限的分布类，则方差 $\theta=E_F(X_1-E_FX_1)^2$ 是对 $\mathscr{F}$ 自由度为 2 的可估参数. 对称核为

$$\begin{aligned}h(X_1,X_2)&=\frac{1}{2}[(X_1^2-X_1X_2)+(X_2^2-X_1X_2)]\\&=\frac{1}{2}(X_1-X_2)^2,\end{aligned}$$

相应的 U 统计量为

$$\begin{aligned}U(X_1,\cdots,X_n)&=\frac{1}{\binom{n}{2}}\sum_{\beta_1<\beta_2}\frac{1}{2}(X_{\beta_1}-X_{\beta_2})^2\\&=\frac{1}{n(n-1)}\sum_{\beta_1<\beta_2}(X_{\beta_1}^2+X_{\beta_2}^2-2X_{\beta_1}X_{\beta_2})\\&=\frac{1}{n(n-1)}\Big[\frac{1}{2}\sum_{\beta_1\neq\beta_2}(X_{\beta_1}^2+X_{\beta_2}^2)-\sum_{\beta_1\neq\beta_2}X_{\beta_1}X_{\beta_2}\Big]\\&=\frac{1}{n(n-1)}\Big[\frac{1}{2}\sum_{\beta_1=1}^{n}\sum_{\beta_2=1}^{n}(X_{\beta_1}^2+X_{\beta_2}^2)\\&\quad-\frac{1}{2}\sum_{\beta_1=1}^{n}(X_{\beta_1}^2+X_{\beta_1}^2)-\sum_{\beta_1\neq\beta_2}X_{\beta_1}X_{\beta_2}\Big]\\&=\frac{1}{n(n-1)}\Big[n\sum_{i=1}^{n}X_i^2-\Big(\sum_{i=1}^{n}X_i\Big)^2\Big]\end{aligned}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

例 3 估一随机变量 X 取正值的概率 $P\{X>0\}$，则可取

$$h(X_1) = \psi(X_1) = \begin{cases} 1, & \text{当 } X_1 > 0, \\ 0, & \text{当 } X_1 \leqslant 0. \end{cases}$$

对一切分布函数 $F(x)$ 均有

$$E_F h(X_1) = E_F \psi(X_1) = P\{X_1 > 0\} = 1 - F(0),$$

参数 $\theta = P\{X>0\}$ 对应的 U 统计量为

$$U(X_1, \cdots, X_n) = \frac{1}{n}\sum_{i=1}^{n}\psi(X_i)$$

$$= \frac{1}{n}(X_1, \cdots, X_n \text{ 中为正的个数}).$$

通过简单的运算可得 U 统计量方差的一般表达式：

$$\operatorname{var}(U) = E\left[\frac{1}{\binom{n}{r}}\sum_{(\beta_1,\cdots,\beta_r)}(h(X_{\beta_1},\cdots,X_{\beta_r}) - \theta)\right]^2$$

$$= \frac{1}{\binom{n}{r}^2}\sum_{(\beta_1,\cdots,\beta_r)}\sum_{(\alpha_1,\cdots,\alpha_r)}E[(h(X_{\beta_1},\cdots,X_{\beta_r}) - \theta)$$

$$\cdot (h(X_{\alpha_1},\cdots,X_{\alpha_r}) - \theta)]$$

$$= \frac{1}{\binom{n}{r}^2}\sum_{(\beta_1,\cdots,\beta_r)}\sum_{(\alpha_1,\cdots,\alpha_r)}\operatorname{cov}[h(X_{\beta_1},\cdots,X_{\beta_r}), h(X_{\alpha_1},\cdots,X_{\alpha_r})]$$

$$= \frac{1}{\binom{n}{r}^2}\sum_{c=0}^{r}\binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}\zeta_c$$

$$= \frac{1}{\binom{n}{r}}\sum_{c=1}^{r}\binom{r}{c}\binom{n-r}{r-c}\zeta_c, \tag{2.1.2}$$

其中 $\zeta_c = \operatorname{cov}[h(X_1,\cdots,X_c,X_{c+1},\cdots,X_r), h(X_1,\cdots,X_c,X_{r+1},\cdots,$

$X_{2r-c})]$，即两个涉及的样本中恰有 c 个相同的两个核的协方差. 特别有 $\zeta_0=0,\zeta_r=\mathrm{var}[h(X_1,\cdots,X_r)]$.

对 $c=0,1,\cdots,r$ 有

$$0\leqslant\zeta_c\leqslant\zeta_r, \tag{2.1.3}$$

这是由于

$$\begin{aligned}\zeta_c= & \operatorname{cov}[h(X_1,\cdots,X_c,X_{c+1},\cdots,X_r),\\ & \qquad h(X_1,\cdots,X_c,X_{r+1},\cdots,X_{2r-c})]\\ \leqslant & \{\mathrm{var}[h(X_1,\cdots,X_c,X_{c+1},\cdots,X_r)]\\ & \quad\cdot\mathrm{var}[h(X_1,\cdots,X_c,X_{r+1},\cdots,X_{2r-c})]\}^{1/2}\\ = & \{\mathrm{var}[h(X_1,\cdots,X_r)]\cdot\mathrm{var}[h(X_1,\cdots,X_r)]\}^{1/2}\\ = & \mathrm{var}[h(X_1,\cdots,X_r)]=\zeta_r.\end{aligned}$$

又

$$\begin{aligned}\zeta_c= & E\{[h(X_1,\cdots,X_c,X_{c+1},\cdots,X_r)-\theta]\\ & \quad\cdot[h(X_1,\cdots,X_c,X_{r+1},\cdots,X_{2r-c})-\theta]\}\\ = & E_{X_1,\cdots,X_c}\{E_{X_{c+1},\cdots,X_r}[h(X_1,\cdots,X_c,X_{c+1},\cdots,X_r)-\theta]\\ & \quad\cdot E_{X_{r+1},\cdots,X_{2r-c}}[h(X_1,\cdots,X_c,X_{r+1},\cdots,X_{2r-c})-\theta]\}\\ = & E_{X_1,\cdots,X_c}\{E_{X_{c+1},\cdots,X_r}[h(X_1,\cdots,X_c,X_{c+1},\cdots,X_r)-\theta]\}^2\\ \geqslant & 0.\end{aligned}$$

利用上述 U 统计量方差的一般表达式计算其方差.

例 4 求总体期望值的 U 统计量$\overline{X}$的方差. 利用上述 U 统计量方差的一般表达式，此时，$r=1,h(X)=X,\zeta_1=\mathrm{var}(X_1)=\sigma^2$(总体方差)，

$$\mathrm{var}(\overline{X})=\frac{1}{\binom{n}{1}}\binom{1}{1}\binom{n-1}{0}\zeta_1=\frac{\sigma^2}{n}.$$

例 5 求总体方差的 U 统量 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 的方差.

利用上述 U 统计量方差的一般表达式，此时，$r=2$，

$$h(X_1,X_2)=\frac{1}{2}(X_1-X_2)^2,$$

$$\begin{aligned}
\zeta_1 &= \operatorname{cov}\left[\frac{1}{2}(X_1-X_2)^2, \frac{1}{2}(X_1-X_3)^2\right] \\
&= \frac{1}{4}E(X_1-X_2)^2(X_1-X_3)^2-[E(X_1-E(X_1))^2]^2 \\
&= \frac{1}{4}E\{[(X_1-E(X_1))-(X_2-E(X_2))]^2 \\
&\quad \cdot [(X_1-E(X_1))-(X_3-E(X_3))]^2\} \\
&\quad -[E(X_1-E(X_1))^2]^2 \\
&= \frac{1}{4}\{E(X_1-E(X_1))^4+3[E(X_1-E(X_1))^2]^2\} \\
&\quad -[E(X_1-E(X_1))^2]^2 \\
&= \frac{1}{4}\{E[(X_1-E(X_1))^2]^2-[E(X_1-E(X_1))^2]^2\} \\
&= \frac{1}{4}\operatorname{var}[(X_1-E(X_1))^2],
\end{aligned}$$

$$\begin{aligned}
\zeta_2 &= \operatorname{var}\left[\frac{1}{2}(X_1-X_2)^2\right] \\
&= \frac{1}{4}E(X_1-X_2)^4-[E(X_1-E(X_1))^2]^2 \\
&= \frac{1}{4}E[(X_1-E(X_1))-(X_2-E(X_2))]^4 \\
&\quad -[E(X_1-E(X_1))^2]^2 \\
&= \frac{1}{2}\{E(X_1-E(X_1))^4+3[E(X_1-E(X_1))^2]^2\} \\
&\quad -[E(X_1-E(X_1))^2]^2 \\
&= \frac{1}{2}\{E(X_1-E(X_1))^4-[E(X_1-E(X_1))^2]^2\} \\
&\quad +[E(X_1-E(X_1))^2]^2 \\
&= \frac{1}{2}\operatorname{var}[(X_1-E(X_1))^2]+[E(X_1-E(X_1))^2]^2;
\end{aligned}$$

$$\operatorname{var}(S^2)=\frac{1}{\binom{n}{2}}[2(n-2)\zeta_1+\zeta_2]$$

$$=\frac{2}{n(n-1)}[E(X_1-E(X_1))^2]^2+\frac{1}{n}\mathrm{var}[(X_1-E(X_1))^2]$$

$$=\frac{2}{n(n-1)}[E(X_1-E(X_1))^2]^2+\frac{1}{n}E(X_1-E(X_1))^4$$

$$-\frac{1}{n}[E(X_1-E(X_1))^2]^2$$

$$=\frac{1}{n}E(X_1-E(X_1))^4-\frac{n-3}{n(n-1)}[E(X_1-E(X_1))^2]^2.$$

例 6 求总体分布取正值的概率 $\theta=1-F(0)$ 的 U 统计量 $B=\frac{1}{n}\sum_{i=1}^{n}\psi(X_i)$ 的方差.

利用上述 U 统计量方差的一般表达式,此时,$r=1,h(X)=\psi(X),\zeta_1=\mathrm{var}[\psi(X)]=F(0)(1-F(0))$,

$$\mathrm{var}(B)=\frac{1}{n}F(0)(1-F(0)).$$

§2.2 一样本 U 统计量的渐近分布

本节讨论 U 统计量的大样本性质.

定理 2.1 设随机变量 $X_1,\cdots,X_n$ 是来自 $F(x)$ 的简单随机样本. $F(x)$ 有未知参数 θ,θ 是 r 可估参数. $U(X_1,\cdots,X_n)$ 是 θ 的 U 统计量,其核为 $h(X_1,\cdots,X_r)$,有

$$E[h(X_1,\cdots,X_r)]^2<\infty,$$

则

$$\lim_{n\to\infty} n\mathrm{var}[U(X_1,\cdots,X_n)]=r^2\zeta_1. \tag{2.2.1}$$

证明 由定理条件知 $\zeta_r=\mathrm{var}[h(X_1,\cdots,X_r)]<\infty$. 而对(2.1.2)式中的 ζ_c,有 $0\leqslant\zeta_c\leqslant\zeta_r$ $(c=1,\cdots,r)$,所以 ζ_c 均有限. 由 U 统计量方差的一般表达式,

$$n\mathrm{var}(U)=\sum_{c=1}^{r}\frac{n\binom{r}{c}\binom{n-r}{r-c}}{\binom{n}{r}}\zeta_c$$

$$= \sum_{c=1}^{r}\binom{r}{c}\frac{r!}{(r-c)!}\frac{n\cdot(n-r)\cdots(n-2r+c+1)}{n\cdot(n-1)\cdots(n-r+1)}\zeta_c.$$

上式中，$\binom{r}{c}\frac{r!}{(r-c)!}$与 n 无关是常数，而当 $n\to\infty$时，若 $c=2,\cdots,r$，则

$$\frac{n\cdot(n-r)\cdots(n-2r+c+1)}{n(n-1)\cdots(n-r+1)}\to 0;$$

若 $c=1$，则

$$\frac{n(n-r)\cdots(n-2r+c+1)}{n(n-1)\cdots(n-r+1)}\to 1.$$

所以

$$\lim_{n\to\infty} n\operatorname{var}(U) = \binom{r}{1}\frac{r!}{(r-1)!}\zeta_1 = r^2\zeta_1.$$

证毕.

由定理知当 $E[h(X_1,\cdots,X_r)]^2<\infty$时，$U(X_1,\cdots,X_n)$均方收敛到$\theta$. 从而 U 是 θ 的相合估计.

以下进一步证明 U 统计量是渐近正态的.

定义 2.1 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布为 $F(t)$，假定统计量 $W(X_1,\cdots,X_n)$对变元是对称的，且 $E(W)=0$. 又随机变量的集合

$$\mathscr{V} = \left\{V\,|\,V = \sum_{i=1}^{n}K(X_i), K(\cdot)\text{ 是一实函数}\right\},$$

则 W 在 $\mathscr{V}$ 上的投影定义为

$$V^* = \sum_{i=1}^{n}K^*(X_i), \tag{2.2.2}$$

其中 $K^*(t)=E\{W\,|\,X_1=t\}$.

定理 2.2 设统计量 $W(X_1,\cdots,X_n)$是独立同分布的随机变量 $X_1,\cdots,X_n$ 的对称统计量，有 $E(W)=0$. 又

$$V^* = \sum_{i=1}^{n}K^*(X_i)$$

是 W 在 $\mathscr{V}$ 上的投影，$K^*(t)=E\{W|X_1=t\}$，则对任一 $V\in\mathscr{V}$，有

$$E(W-V^*)^2\leqslant E(W-V)^2.$$

证明 不等式右端的项

$$\begin{aligned}E(W-V)^2&=E[(W-V^*)+(V^*-V)]^2\\&=E(W-V^*)^2+E(V^*-V)^2\\&\quad+2E(W-V^*)(V^*-V).\end{aligned}$$

因为 $V^*=\sum_{i=1}^{n}K^*(X_i)$，$V=\sum_{i=1}^{n}K(X_i)$，故交叉项

$$E(W-V^*)(V^*-V)=\sum_{i=1}^{n}E(W-V^*)(K^*(X_i)-K(X_i)).$$

而

$$\begin{aligned}&E(W-V^*)(K^*(X_i)-K(X_i))\\&\quad=E_{X_i}E[(W-V^*)(K^*(X_i)-K(X_i))|X_i]\\&\quad=E_{X_i}\{(K^*(X_i)-K(X_i))E[(W-V^*)|X_i]\}.\end{aligned}$$

但上式中的

$$\begin{aligned}E[(W-V^*)|X_i]&=E(W|X_i)-\sum_{j=1}^{n}E(K^*(X_j)|X_i)\\&=K^*(X_i)-K^*(X_i)-(n-1)E_{X_j}(K^*(X_j))\\&=-(n-1)E_{X_j}E(W|X_j)\\&=-(n-1)E(W)=0,\end{aligned}$$

从而交叉项

$$E(W-V^*)(V^*-V)=0,$$

$$E(W-V)^2=E(W-V^*)^2+E(V^*-V)^2.$$

因此

$$E(W-V^*)^2\leqslant E(W-V)^2.$$

证毕.

上述定理说明当距离为均方距离时，V^* 是通常意义下 W 在 $\mathscr{V}$ 上的投影.

例 1 总体的方差 $\theta=\mathrm{var}(X)$ 的 U 统计量为 $U(X_1,\cdots,X_n)=$

$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$,求它在 $\mathscr{V}$ 上的投影.

$$K^*(t)=E[S^2-\theta|X_1=t]$$
$$=E\left[\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\overline{X}^2\right)-\theta|X_1=t\right]$$
$$=E\left[\frac{1}{n(n-1)}\left\{n\left(\sum_{i=1}^{n}X_i^2\right)-\left(\sum_{i=1}^{n}X_i^2+2\sum_{i<j}\sum X_iX_j\right)\right\}\right.$$
$$\left.-\theta|X_1=t\right]$$
$$=\frac{1}{n(n-1)}E\left[(n-1)\left(t^2+\sum_{i=2}^{n}X_i^2\right)-2t\sum_{i=2}^{n}X_i\right.$$
$$\left.-2\sum_{2\leqslant i<j\leqslant n}\sum X_iX_j|X_1=t\right]-\theta$$
$$=\frac{1}{n(n-1)}\{(n-1)[t^2+(n-1)(\theta+\mu^2)$$
$$-2(n-1)t\mu-(n-1)(n-2)\mu^2\}-\theta$$
$$=\frac{1}{n}[(t-\mu)^2-\theta],$$

其中 $\mu=E(X)$是总体的期望.因此 $S^2-\theta$ 在 $\mathscr{V}$ 上的投影为

$$V^*=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\theta.$$

对一般的 U 统计量,若其对称核为 $h(X_1,\cdots,X_r)$,记

$$E[h(X_1,\cdots,X_r)|X_1=t]=h^*(t). \tag{2.2.3}$$

对 $h^*(\cdot)$有

$$E_{X_i}h^*(X_i)=E[h(X_1,\cdots,X_r)]=\theta, \tag{2.2.4}$$

$$\mathrm{var}_{X_i}h^*(X_i)=E_{X_i}[h^*(X_i)]^2-\theta^2$$
$$=E_{X_1}[h^*(X_1)]^2-\theta^2$$
$$=E_{X_1}\{E[h(X_1,X_2,\cdots,X_r)|X_1]$$
$$\cdot E[h(X_1,X_{r+1},\cdots,X_{2r-1})|X_1]\}-\theta^2$$

$$= E_{X_1}\{E[h(X_1,X_2,\cdots,X_r)$$
$$\cdot h(X_1,X_{r+1},\cdots,X_{2r-1})|X_1]\} - \theta^2$$
$$= E\{h(X_1,X_2,\cdots,X_r)$$
$$\cdot h(X_1,X_{r+1},\cdots,X_{2r-1})\} - \theta^2$$
$$= \zeta_1, \qquad (2.2.5)$$

这里 E_{X_i}, var_{X_i}分别表示根据 X_i 的分布求期望,方差.

定理 2.3 若$U(X_1,\cdots,X_n)$是自由度 r 的可估参数 θ 的 U 统计量,其核为 $h(X_1,\cdots,X_r)$,则 $U(X_1,\cdots,X_n)-\theta$ 在 $\mathscr{V}$ 上的投影为

$$V^* = \frac{r}{n}\sum_{i=1}^{n}[h^*(X_i) - \theta], \qquad (2.2.6)$$

其中 $h^*(t)=E[h(X_1,\cdots,X_r)|X_1=t]$.

证明 根据定义 2.1,(2.2.2)式中的

$$K^*(t) = E\{U(X_1,\cdots,X_n) - \theta|X_1 = t\}$$
$$= E\left\{\frac{1}{\binom{n}{r}}\sum_{(\beta_1,\cdots,\beta_r)}h(X_{\beta_1},\cdots,X_{\beta_r}) - \theta|X_1 = t\right\}$$
$$= \frac{1}{\binom{n}{r}}\left\{\binom{n-1}{r}\theta + \binom{n-1}{r-1}h^*(t)\right\} - \theta$$
$$= \frac{r}{n}h^*(t) + \frac{n-r}{n}\theta - \theta = \frac{r}{n}[h^*(t) - \theta],$$

故投影 $V^* = \frac{r}{n}\sum_{i=1}^{n}[h^*(X_i)-\theta]$. 证毕.

例 2 利用上述定理求总体方差的 U 统计量

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

的投影.

S^2 的核为

$$h(X_1,X_2) = \frac{1}{2}(X_1 - X_2)^2.$$

(2.2.6)式中的

$$\begin{aligned}h^*(t) &= E[h(X_1,X_2)\,|\,X_1=t] = E(h(t,X_2))\\ &= \frac{1}{2}E(t-X_2)^2 = \frac{1}{2}(t^2-2t\mu+E(X_2^2))\\ &= \frac{1}{2}(t^2-2t\mu+\mu^2)+\frac{1}{2}E(X_2-\mu)^2\\ &= \frac{1}{2}(t-\mu)^2+\frac{1}{2}\theta.\end{aligned}$$

故根据(2.2.6)式,$S^2-\theta$ 的投影为

$$\begin{aligned}V^* &= \frac{2}{n}\sum_{i=1}^{n}\left[\frac{1}{2}(X_i-\mu)^2+\frac{1}{2}\theta-\theta\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}[(X_i-\mu)^2-\theta],\end{aligned} \tag{2.2.7}$$

其中 μ,θ 分别为总体的期望和方差.

例 3 求概率 $\theta=P\{X_1+X_2>0\}$ 的 U 统计量

$$U(X_1,\cdots,X_n)=\frac{1}{\binom{n}{2}}\sum_{i<j}\sum\psi(X_i+X_j)$$

的投影.

(2.2.6)式中的

$$\begin{aligned}h^*(t) &= E\{\psi(X_1+X_2)\,|\,X_1=t\} = E\{\psi(t+X_2)\}\\ &= P\{t+X_2>0\} = P\{X_2>-t\} = 1-F(-t).\end{aligned}$$

故根据(2.2.6)式,$U(X_1,\cdots,X_n)-\theta$ 在 $\mathscr{V}$ 上的投影为

$$V^* = \frac{2}{n}\sum_{i=1}^{n}[1-F(-X_i)-\theta], \tag{2.2.8}$$

其中 $F(t)$ 是总体的分布函数. 当 $F(t)$ 关于 0 点对称时,则有

$$V^* = \frac{2}{n}\sum_{i=1}^{n}[F(X_i)-\theta]. \tag{2.2.9}$$

定理 2.4(Hoeffding 定理) 设随机变量 $X_1,\cdots,X_n$ 是来自某总体的简单随机样本,θ 是总体的 r 可估参数,其对称核为

$h(X_1,\cdots,X_r)$. 若

$$E[h(X_1,\cdots,X_r)]^2 < \infty,$$

则对 U 统计量

$$U(X_1,\cdots,X_n) = \frac{1}{\binom{n}{r}} \sum_{(\beta_1,\cdots,\beta_r)} h(X_{\beta_1},\cdots,X_{\beta_r})$$

当 $\zeta_1 = \text{cov}[h(X_1,X_2,\cdots,X_r), h(X_1,X_{r+1},\cdots,X_{2r-1})] > 0$ 时，统计量

$$\sqrt{n}\,[U(X_1,\cdots,X_n) - \theta]$$

的极限分布($n\to\infty$)为 $N(0,r^2\zeta_1)$.

证明 记 $U_n = U(X_1,\cdots,X_n)$. 由定理 2.3，$U_n - \theta$ 在 $\mathscr{V}$ 上的投影为

$$V_n^* = \frac{r}{n}\sum_{i=1}^{n}[h^*(X_i) - \theta],$$

由(2.2.4)，(2.2.5)和(2.1.3)知 $E(h^*(X_i)) = \theta$，$\text{var}h^*(X_i) = \zeta_1 < \zeta_r = E[h(X_1,\cdots,X_r) - \theta]^2$，从而 $E(V_n^*) = 0$，

$$\text{var}(V_n^*) = \left(\frac{r}{n}\right)^2 \sum_{i=1}^{n}\text{var}[h^*(X_i)] = \frac{r^2}{n}\zeta_1,$$

其中 V_n^* 是二阶矩存在的独立同分布随机变量的和，所以 $\sqrt{n}\,V_n^*$ 有极限分布 $N(0,r^2\zeta_1)$. 我们若能证明

$$E[\sqrt{n}\,(U_n - \theta) - \sqrt{n}\,V_n^*]^2 \to 0,$$

则由 Slutsky 定理即可得本定理结论. 下面来证明上列关系式. 由于

$$\begin{aligned} &nE[(U_n - \theta) - V_n^*]^2 \\ &\quad = nE(U_n - \theta)^2 - nE(V_n^*)^2 - 2nE[(U_n - \theta)V_n^*], \end{aligned}$$

由本节定理 2.1 知 $nE(U_n-\theta)^2 \to r^2\zeta_1$，又 $nE(V_n^*)^2 = r^2\zeta_1$，故只需再证

$$nE[(U_n - \theta)V_n^*] \to r^2\zeta_1.$$

而由 U_n 和 V_n^* 的定义，有

$$nE[(U_n-\theta)V_n^*]$$

$$=nE\left\{\frac{1}{\binom{n}{r}}\sum_{(\beta_1,\cdots,\beta_r)}[h(X_{\beta_1},\cdots,X_{\beta_r})-\theta]\cdot\frac{r}{n}\sum_{i=1}^{n}[h^*(X_i)-\theta]\right\}$$

$$=\frac{r}{\binom{n}{r}}\sum_{i=1}^{n}\sum_{(\beta_1,\cdots,\beta_r)}E\{[h(X_{\beta_1},\cdots,X_{\beta_r})-\theta][h^*(X_i)-\theta]\}.$$

当 $i\overline{\in}\{\beta_1,\cdots,\beta_r\}$时，

$$E\{[h(X_{\beta_1},\cdots,X_{\beta_r})-\theta][h^*(X_i)-\theta]\}=0;$$

当 $i\in\{\beta_1,\cdots,\beta_r\}$时，

$$E\{[h(X_{\beta_1},\cdots,X_{\beta_r})-\theta][h^*(X_i)-\theta]\}$$

$$=E_{X_i}E\{[h(X_{\beta_1},\cdots,X_{\beta_r})-\theta][h^*(X_i)-\theta]|X_i\}$$

$$=E_{X_i}\{[h^*(X_i)-\theta]E[h(X_{\beta_1},\cdots,X_{\beta_r})-\theta|X_i]\}$$

$$=E_{X_i}[h^*(X_i)-\theta]^2=\zeta_1.$$

所以

$$nE[(U_n-\theta)V_n^*]=\frac{r}{\binom{n}{r}}\sum_{i=1}^{n}\binom{n-1}{r-1}\zeta_1=r^2\zeta_1.$$

从而

$$nE[(U_n-\theta)-V_n^*]^2\to r^2\zeta_1+r^2\zeta_1-2r^2\zeta_1=0.$$

故有定理结论成立. 证毕.

例 4 考虑计数统计量 $B=\sum_{i=1}^{n}\psi(X_i)$(参见第一章第 2 节). B/n 是自由度为 1 的可估参数 $P\{X>0\}$的 U 统计量，

$$\zeta_1=\mathrm{var}[\psi(X_1)]=P\{X>0\}\cdot[1-P\{X>0\}].$$

所以有$\sqrt{n}\left(\frac{B}{n}-P\{X>0\}\right)$渐近正态分布 $N(0,\zeta_1)$，即

$$\frac{B-E(B)}{\sqrt{\text{var}B}}=\frac{n\left(\frac{1}{n}B-P\{X>0\}\right)}{\sqrt{n\text{var}[\psi(X_1)]}}=\frac{\sqrt{n}\left(\frac{1}{n}B-P\{X>0\}\right)}{\sqrt{\zeta_1}}$$

渐近正态分布 $N(0,1)$.

例 5 考虑符号秩统计量 $W^+=\sum_{i=1}^{n}\psi(X_i)R_i^+$（参见第一章第 4 节）. 由第一章第 4 节(1.4.6)式知

$$W^+=\sum_{i=1}^{n}\psi(X_i)R_i^+=\sum_{i=1}^{n}\sum_{j\leqslant i}\psi(X_i+X_j)$$

$$=\sum_{i=1}^{n}\psi(X_i)+\sum_{i=1}^{n}\sum_{j<i}\psi(X_i+X_j).$$

而

$U_1=\frac{1}{n}\sum_{i=1}^{n}\psi(X_i)$ 是 $P\{X_1>0\}$ 的 U 统计量；

$U_2=\frac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{j<i}\psi(X_i+X_j)$ 是 $P\{X_1+X_2>0\}$ 的 U 统计量，

故

$$\sqrt{n}\left(\frac{W^+-E(W^+)}{\binom{n}{2}}\right)=\frac{n^{3/2}}{\binom{n}{2}}(U_1-E(U_1))+\sqrt{n}(U_2-E(U_2)).$$

易见

$$\frac{n^{3/2}}{\binom{n}{2}}(U_1-E(U_1))\xrightarrow{P}0.$$

而由定理 2.4，

$$\sqrt{n}(U_2-E(U_2))$$

有渐近正态分布 $N(0,4\zeta_1)$，其中

$$\zeta_1=\text{cov}[\psi(X_1+X_2),\psi(X_1+X_3)]$$
$$=P\{X_1+X_2>0,X_1+X_3>0\}-[P\{X_1+X_2>0\}]^2.$$

因而

$$\frac{\sqrt{n}}{\binom{n}{2}}(W^+ - E(W^+))$$

有渐近正态分布 $N(0,4\zeta_1)$. 当总体分布连续,关于 0 点对称时,可以算得

$$\begin{aligned}
P\{X_1 + X_2 > 0\} &= \int_{x_1+x_2>0} \mathrm{d}F(x_1)\mathrm{d}F(x_2) \\
&= \int_{-\infty}^{+\infty}\left[\int_{-x_2}^{\infty}\mathrm{d}F(x_1)\right]\mathrm{d}F(x_2) \\
&= \int_{-\infty}^{+\infty}[1 - F(-x_2)]\mathrm{d}F(x_2) \\
&= \int_{-\infty}^{+\infty}F(x_2)\mathrm{d}F(x_2) = \frac{1}{2}[F(x_2)]^2\bigg|_{-\infty}^{+\infty} = \frac{1}{2},
\end{aligned}$$

$$\begin{aligned}
&P\{X_1 + X_2 > 0, X_1 + X_3 > 0\} \\
&= \int_{-\infty}^{+\infty}\left[\left(\int_{-x_1}^{+\infty}\mathrm{d}F(x_2)\right)\left(\int_{-x_1}^{+\infty}\mathrm{d}F(x_3)\right)\right]\mathrm{d}F(x_1) \\
&= \frac{1}{3}.
\end{aligned}$$

所以有 $4\zeta_1 = \frac{1}{3}$. 故此时,

$$\frac{\sqrt{n}}{\binom{n}{2}}\cdot\frac{(W^+ - E(W^+))}{\sqrt{\frac{1}{3}}}$$

有渐近正态分布 $N(0,1)$. 而

$$\begin{aligned}
&\sqrt{\mathrm{var}W^+}\bigg/\left[\binom{n}{2}\cdot\sqrt{\frac{1}{3n}}\right] \\
&\qquad = \sqrt{\frac{n(n+1)(2n+1)}{24}}\bigg/\left[\frac{n(n-1)}{2}\sqrt{\frac{1}{3n}}\right] \to 1,
\end{aligned}$$

所以此时亦有

$$\frac{W^+ - E(W^+)}{\sqrt{\operatorname{var} W^+}}$$ 有渐近正态分布 $N(0,1)$.

§2.3 二样本 U 统计量的渐近分布

定义 3.1 有两个独立样本：$X_1,\cdots,X_m$ 相互独立同分布，分布为 $F(x)$；$Y_1,\cdots,Y_n$ 相互独立同分布，分布为 $G(y)$. 参数 θ 称为对分布族 $\mathscr{F}=\{(F,G)\}$ 是 (r,s) 可估的，如果存在样本 $X_1,\cdots,X_r$；$Y_1,\cdots,Y_s$ 的函数 $h(X_1,\cdots,X_r;Y_1,\cdots,Y_s)$ 使对一切 $(F,G)\in\mathscr{F}$，有

$$E_{(F,G)}[h(X_1,\cdots,X_r;Y_1,\cdots,Y_s)] = \theta,$$

其中 $E_{(F,G)}h(\cdot)$ 表示统计量 $h(\cdot)$ 在两个总体分布分别为 F,G 时的期望值. 若上述关系成立的最小样本量为 r,s，则 (r,s) 称为**可估参数 θ 的自由度**. $h(\cdot\,;\,\cdot)$ 称为**核**.

我们不妨认为 $h(\cdot\,;\,\cdot)$ 对 $X_1,\cdots,X_r$ 和 $Y_1,\cdots,Y_s$ 分别对称.

定义 3.2 条件同定义 3.1，由对称核 $h(X_1,\cdots,X_r;Y_1,\cdots,Y_s)$ 形成的统计量

$$U(X_1,\cdots,X_m;Y_1,\cdots,Y_n) = \frac{1}{\binom{m}{r}\binom{n}{s}}\sum_{(\alpha_1,\cdots,\alpha_r)}\sum_{(\beta_1,\cdots,\beta_s)} h(X_{\alpha_1},\cdots,X_{\alpha_r};Y_{\beta_1},\cdots,Y_{\beta_s}) \tag{2.3.1}$$

称为**可估参数 θ 的 U 统计量**.

例 1 概率 $\theta=P\{X<Y\}$ 是 $(1,1)$ 可估参数，其对称核为 $\psi(Y_1-X_1)$，对应的 U 统计量为

$$U = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(Y_j - X_i),$$

$\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(Y_j-X_i)$ 是第一章第 3 节中的 Mann-Whitney 统计量.

类似一样本的情形，可求二样本 U 统计量的方差. 记

$$\zeta_{c,d} = \mathrm{cov}[h(X_1,\cdots,X_c,X_{c+1},\cdots,X_r;Y_1,\cdots,Y_d,Y_{d+1},\cdots,Y_s),$$
$$h(X_1,\cdots,X_c,X_{r+1},\cdots,X_{2r-c};Y_1,\cdots,Y_d,Y_{s+1},\cdots,Y_{2s-d})],$$

则二样本 U 统计量的方差有一般表达式：

$$\mathrm{var}\, U = \frac{1}{\binom{m}{r}\binom{n}{s}} \sum_{c=0}^{r}\sum_{d=0}^{s} \binom{r}{c}\binom{m-r}{r-c}\binom{s}{d}\binom{n-s}{s-d}\zeta_{c,d}. \tag{2.3.2}$$

定理 3.1 设统计量 $U(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$ 是 (r,s) 可估参数 θ 的 U 统计量，其对称核为 $h(X_1,\cdots,X_r;Y_1,\cdots,Y_s)$. 若 $E[h(X_1,\cdots,X_r;Y_1,\cdots,Y_s)]^2<\infty$, 又有 $\frac{m}{n+m}\to\lambda$ $(0<\lambda<1)$, 当 $N=n+m\to\infty$, 则有

$$\lim_{N\to\infty} N\mathrm{var}[U(X_1,\cdots,X_m;Y_1,\cdots,Y_n)] = \frac{r^2\zeta_{1,0}}{\lambda} + \frac{s^2\zeta_{0,1}}{1-\lambda}. \tag{2.3.3}$$

证明 类似前一节定理 2.1.

与一样本的情形类似亦可定义投影. 记随机变量集合

$$\mathscr{V} = \left\{ \sum_{i=1}^{m} K_1(X_i) + \sum_{j=1}^{n} K_2(Y_j) \right\},$$

$(U-\theta)$ 到 $\mathscr{V}$ 的**投影**定义为

$$V^* = \sum_{i=1}^{m} K_1^*(X_i) + \sum_{j=1}^{n} K_2^*(Y_j), \tag{2.3.4}$$

其中

$$K_1^*(t) = E[U-\theta|X_1=t],$$
$$K_2^*(t) = E[U-\theta|Y_1=t].$$

定理 3.2(Lehmann-Hoeffding 定理) 设随机变量 $X_1,\cdots,X_m$ 相互独立同分布，分布为 $F(x)$；随机变量 $Y_1,\cdots,Y_n$ 相互独立同分布，分布为 $G(y)$，统计量 $U(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$ 是 (r,s) 可估参数 θ 的 U 统计量，其对称核为 $h(X_1,\cdots,X_r;Y_1,\cdots,Y_s)$. 若

$$E[h(X_1,\cdots,X_r;Y_1,\cdots,Y_s)]^2 < \infty,$$

又当 $N=m+n\to\infty$ 时，$\frac{m}{N}\to\lambda\ (0<\lambda<1)$，则当

$$\frac{r^2\zeta_{1,0}}{\lambda}+\frac{s^2\zeta_{0,1}}{1-\lambda}>0 \tag{2.3.5}$$

时，$\sqrt{N}\,[U(X_1,\cdots,X_m;Y_1,\cdots,Y_n)-\theta]$ 有渐近正态分布 $N\left(0,\frac{r^2\zeta_{1,0}}{\lambda}+\frac{s^2\zeta_{0,1}}{1-\lambda}\right)$.

证明 类似前节定理 2.4.

例 2 考虑 Wilcoxon 秩和统计量 $W=\sum_{i=1}^{n}R_i$(参见第一章第 3 节).

由第一章第 3 节(1.3.12)式知

$$W=\frac{n(n+1)}{2}+\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(Y_j-X_i).$$

而统计量

$$U=\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(Y_j-X_i)$$

是(1,1)可估参数 $\theta=P\{X<Y\}$ 的 U 统计量，其对称核为 $h(X_1,Y_1)=\psi(Y_1-X_1)$. 对应的

$$\begin{aligned}\zeta_{1,0}&=\text{cov}[\psi(Y_1-X_1),\psi(Y_2-X_1)]\\&=P\{X_1<Y_1,X_1<Y_2\}-[P\{X_1<Y_1\}]^2,\\ \zeta_{0,1}&=\text{cov}[\psi(Y_1-X_1),\psi(Y_1-X_2)]\\&=P\{X_1<Y_1,X_2<Y_1\}-[P\{X_1<Y_1\}]^2,\end{aligned}$$

在 H_0: $X_1,\cdots,X_m;Y_1,\cdots,Y_n$ 独立同分布，分布函数为 $F(x)$，$F(x)$ 连续时，可算出

$$P\{X_1<Y_1\}=\frac{1}{2},$$

$$P\{X_1<Y_1,X_1<Y_2\}=\frac{1}{3},$$

$$P\{X_1<Y_1,X_2<Y_1\}=\frac{1}{3}.$$

此时，$\zeta_{1,0}=\zeta_{0,1}=\frac{1}{12}$. 故在 H_0 下，

$$\sqrt{N}\left(U-\frac{1}{2}\right)\text{有渐近正态分布 } N\left(0,\frac{1}{12\lambda}+\frac{1}{12(1-\lambda)}\right).$$

而 Wilcoxon 秩和统计量

$$W=mnU+\frac{n(n+1)}{2},$$

$$E(W)=\frac{n(N+1)}{2},\quad \mathrm{var}W=\frac{mn(N+1)}{12}.$$

易见

$$\mathrm{var}W\Big/\left[\frac{m^2n^2}{N}\cdot\frac{1}{12\lambda(1-\lambda)}\right]\to 1\quad (N\to\infty).\quad (2.3.6)$$

所以

$$\frac{W-E(W)}{\sqrt{\mathrm{var}W}}\text{有渐近正态分布 } N(0,1).$$

习　题

1. 设随机变量 $X_1,\cdots,X_n$ 是来自总体 $F(x)$ 的样本，试对下列参数确定(i) 参数可估的自由度；(ii) 对称核 $h(\cdot)$；(iii) U 统计量；并指明(iv) 适应的分布族 $\mathscr{F}$. 这些参数为

(a) $P\{|X_1|>1\}$;

(b) $P\{X_1+X_2+X_3>0\}$;

(c) $E(X_1-\mu)^3$，其中 μ 为 $F(x)$ 的期望；

(d) $E(X_1-X_2)^4$.

2. 考虑参数 $\theta=P\{X_1+X_2>0\}$，其中随机变量 X_1,X_2 相互独立同分布，有连续分布 $F(x)$. 定义

$$h(x)=1-F(-x).$$

说明 $E(h(X_1))=\theta$. 并请回答：$h(X_1)$是对称核吗？为什么？

3. 设统计量 U_1 和 U_2 是同一参数 θ 的 U 统计量，如果分布族 $\mathscr{F}$ 包括全部连续分布，说明

$$P\{U_1=U_2\}=1.$$

4. 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，令

$$p = P\{X_1 > 0\},$$

则对参数 $\theta=p(1-p)$，

(a) 求 θ 的 U 统计量 U^*；

(b) 求 $\mathrm{var}U^*$；

(c) 当 $n\to\infty$ 时，求 U^* 的渐近方差.

5. 设 U_1 和 U_2 分别是自由度为 r_1 的可估参数 θ_1 和自由度为 r_2 的可估参数 θ_2 的 U 统计量，假定参数 $\theta_1+\theta_2$ 的可估的自由度为 $r=\max(r_1,r_2)$，则统计量 $V=U_1+U_2$ 是 $\theta_1+\theta_2$ 的 U 统计量，说明使

$$\sqrt{n}\,(V-\theta_1-\theta_2)$$

有均值为 0 的正态极限分布的条件，并确定极限分布的方差.

6. 求二样本问题参数 $\theta=\mathrm{var}(X)+\mathrm{var}(Y)$ 的 U 统计量 U^*，并确定 $\sqrt{N}(U^*-\theta)$ 的渐近方差.

7. 考虑二样本尺度参数检验，样本为

$X_1,\cdots,X_m$ 相互独立同分布，分布为 $F(x-\theta_X)$；

$Y_1,\cdots,Y_n$ 相互独立同分布，分布为 $F\left(\dfrac{x-\theta_Y}{\eta}\right)$，

$F(t)$ 连续且关于 0 点对称，θ_X,θ_Y 已知，检验假设

$$H_0:\eta=1,\qquad H_1:\eta>1.$$

令

$$X_i^*=|X_i-\theta_X|,\quad i=1,\cdots,m;$$
$$Y_j^*=|Y_j-\theta_Y|,\quad j=1,\cdots,n.$$

由参数 $\theta=P\{X_1^*<Y_1^*\}$ 的 U 统计量 U^* 确定一个上述假设问题的检验，说明这个检验是非参数 d-free 的. 并求 U^* 的期望和 $\sqrt{N}(U^*-\theta)$ 的渐近方差的表达式(用 $F(x)$ 及 η 表示).

8. 证明定理 3.1.

9. 证明定理 3.2.

第三章　线性秩统计量

§3.1　线性秩统计量的定义

定义 1.1　设 $X_1,\cdots,X_N$ 为 N 个随机变量，其对应的秩向量记为 $R=(R_1,\cdots,R_N)$. 又 $a(1),\cdots,a(N)$ 和 $c(1),\cdots,c(N)$ 是两组常数，组内的 N 个数不全相等. 定义统计量

$$S=\sum_{i=1}^{N}c(i)a(R_i),\qquad(3.1.1)$$

S 称为 R 的**线性秩统计量**，$a(1),\cdots,a(N)$ 被称为**分值**(score)，$c(1),\cdots,c(N)$ 被称为**回归常数**.

例 1　二样本问题：随机变量 $X_1,\cdots,X_m$ 相互独立同分布，分布为 $F(x)$；随机变量 $Y_1,\cdots,Y_n$ 相互独立同分布，分布为 $G(y)$. 混合样本 $X_1,\cdots,X_m,Y_1,\cdots,Y_n$ 对应的秩向量记为

$$R=(Q_1,\cdots,Q_m,R_1,\cdots,R_n).$$

取两组常数

$$c(i)=\begin{cases}0, & 当\ i=1,\cdots,m\ 时,\\ 1, & 当\ i=m+1,\cdots,N,N=m+n\ 时;\end{cases}$$

$$a(i)=i,\quad i=1,\cdots,N,$$

则线性秩统计量

$$\begin{aligned}S&=\sum_{i=1}^{m}c(i)a(Q_i)+\sum_{i=m+1}^{N}c(i)a(R_{i-m})\\&=\sum_{i=1}^{n}R_i.\end{aligned}$$

若取两组常数为

$$c(i)=\begin{cases}0, & 当\ i=1,\cdots,m\ 时,\\ 1, & 当\ i=m+1,\cdots,N\ 时;\end{cases}$$

$$a(i)=\begin{cases}0, & 当\ i\leqslant\dfrac{N+1}{2}\ 时,\\ 1, & 当\ i>\dfrac{N+1}{2}\ 时,\end{cases}$$

则线性秩统计量

$$S=\sum_{i=1}^{n}a(R_i)$$

$$=(Y_1,\cdots,Y_n\ 中大于混合样本中位数的个数).$$

上面两个线性秩统计量均可用于二样本位置问题的检验，即检验问题：样本为

$X_1,\cdots,X_m$ 相互独立同分布，分布为 $F(x)$，$F(x)$连续；

$Y_1,\cdots,Y_n$ 相互独立同分布，分布为 $F(x-\Delta)$，Δ 为未知参数.

检验假设

$$H_0:\ \Delta=0,\qquad H_1:\ \Delta>0.$$

若取两组常数为

$$c(i)=\begin{cases}0, & 当\ i=1,\cdots,m\ 时,\\ 1, & 当\ i=m+1,\cdots,N\ 时;\end{cases}$$

$$a(i)=\left(i-\frac{N+1}{2}\right)^2,$$

则线性秩统计量

$$S=\sum_{i=1}^{n}\left(R_i-\frac{N+1}{2}\right)^2. \tag{3.1.2}$$

此线性秩统计量可用于二样本尺度问题的检验，即检验问题：样本为

$X_1,\cdots,X_m$ 相互独立同分布，分布为 $F(x)$，$F(x)$ 连续；

$Y_1,\cdots,Y_n$ 相互独立同分布，分布为 $F\left(\dfrac{x}{\eta}\right)$.

检验假设

$$H_0:\ \eta=1,\qquad H_1:\ \eta>1.$$

例 2 设有 N 个连续随机变量 $X_1,\cdots,X_N$，$X_i\sim F_i(x)$，X_1，

$\cdots, X_N$ 相互独立. 检验假设

H_0: $X_1, \cdots, X_N$ 同分布;

H_1: $F_i(x) \geqslant F_{i+1}(x)$ $(i=1, \cdots, N-1)$ 且不等式对某些点成立.

备择假设 H_1 即是说样本明显地有渐大的趋势, 这时秩向量 $(R_1, \cdots, R_N)$ 应与 $(1, \cdots, N)$ 较一致. 因而一个直观的检验统计量为

$$\hat{\rho} = \frac{\sum_{i=1}^{N}\left(R_i - \frac{N+1}{2}\right)\left(i - \frac{N+1}{2}\right)}{\sqrt{\sum_{i=1}^{N}\left(R_i - \frac{N+1}{2}\right)^2 \cdot \sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2}}, \quad (3.1.3)$$

但

$$\sum_{i=1}^{N}\left(R_i - \frac{N+1}{2}\right)^2 = \sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2$$

$$= \sum_{i=1}^{N} i^2 - N\left(\frac{N+1}{2}\right)^2 = \frac{N(N-1)(N+1)}{12}.$$

所以(3.1.3)式可化简为

$$\hat{\rho} = \frac{12}{N(N-1)(N+1)}\left[\sum_{i=1}^{N} iR_i - N\left(\frac{N+1}{2}\right)^2\right],$$

(3.1.4)

它的等价的检验统计量为

$$S = \sum_{i=1}^{N} iR_i, \quad (3.1.5)$$

即为取 $c(i)=i, a(i)=i$ $(i=1, \cdots, N)$ 的线性秩统计量.

在第一章定理 3.1、定理 3.2 中已指明: 当 $X_1, \cdots, X_N$ 相互独立同分布, 分布连续时, 其秩向量 $R=(R_1, \cdots, R_N)$ 在集合

$$\mathscr{R} = \{(r_1, \cdots, r_N) \mid (r_1, \cdots, r_N) \text{ 是} (1, \cdots, N) \text{ 的排列}\}$$

(3.1.6)

上均匀分布. 边缘分布亦是均匀分布, 即

$$P\{R_i = r\} = \frac{1}{N}, \quad i = 1, \cdots, N, r = 1, \cdots, N;$$

$$P\{R_i = r, R_j = s\}$$

$$= \begin{cases} \dfrac{1}{N(N-1)}, & r \neq s, r, s = 1, \cdots, N, \\ 0, & \text{其他}. \end{cases}$$

$$i \neq j,\ i, j = 1, \cdots, N.$$

利用这些结果可得下列关于线性秩统计量的定理.

定理 1.1 设 $a(1), \cdots, a(N)$ 是 N 个常数. 若秩向量 $R=(R_1, \cdots, R_N)$ 在集合 $\mathscr{R}$ 上均匀分布, 则

$$E[a(R_i)] = \frac{1}{N}\sum_{k=1}^{N} a(k), \quad i = 1, \cdots, N, \tag{3.1.7}$$

$$\mathrm{var}[a(R_i)] = \frac{1}{N}\sum_{k=1}^{N}[a(k) - \overline{a}]^2, \quad i = 1, \cdots, N, \tag{3.1.8}$$

其中 $\overline{a} = \frac{1}{N}\sum\limits_{k=1}^{N} a(k)$,

$$\mathrm{cov}[a(R_i), a(R_j)] = \frac{-1}{N(N-1)}\sum_{k=1}^{N}[a(k) - \overline{a}]^2, \quad i \neq j, \quad i, j = 1, \cdots, N. \tag{3.1.9}$$

证明 直接按定义计算,

$$E[a(R_i)] = \sum_{k=1}^{N} a(k) \cdot P\{R_i = k\} = \overline{a},$$

$$\mathrm{var}[a(R_i)] = \sum_{k=1}^{N}[a(k) - \overline{a}]^2 \cdot P\{R_i = k\}$$

$$= \frac{1}{N}\sum_{k=1}^{N}[a(k) - \overline{a}]^2,$$

$$\mathrm{cov}[a(R_i), a(R_j)]$$

$$= \sum_{k=1}^{N}\sum_{h \neq k}[a(k) - \overline{a}][a(h) - \overline{a}] \cdot P\{R_i = k, R_j = h\}$$

$$= \frac{1}{N(N-1)}\sum_{k=1}^{N}\sum_{h \neq k}[a(k) - \overline{a}][a(h) - \overline{a}]$$

$$= \frac{1}{N(N-1)}\left\{ \sum_{k=1}^{N}\sum_{h=1}^{N}[a(k)-\bar{a}][a(h)-\bar{a}] - \sum_{k=1}^{N}[a(k)-\bar{a}]^2 \right\}$$

$$= \frac{1}{N(N-1)}\left\{ \left(\sum_{k=1}^{N}[a(k)-\bar{a}] \right)^2 - \sum_{k=1}^{N}[a(k)-\bar{a}]^2 \right\}$$

$$= \frac{-1}{N(N-1)}\sum_{k=1}^{N}[a(k)-\bar{a}]^2.$$

证毕.

定理 1.2 若秩向量 $R=(R_1,\cdots,R_N)$在集合 $\mathscr{R}$ 上均匀分布，则线性秩统计量

$$S = \sum_{i=1}^{N} c(i)a(R_i)$$

有

$$E(S) = N\bar{c}\,\bar{a}, \tag{3.1.10}$$

$$\operatorname{var}(S) = \frac{1}{N-1}\left\{ \sum_{i=1}^{N}[a(i)-\bar{a}]^2 \right\}\left\{ \sum_{i=1}^{N}[c(i)-\bar{c}]^2 \right\}, \tag{3.1.11}$$

其中

$$\bar{c} = \frac{1}{N}\sum_{i=1}^{N}c(i), \quad \bar{a} = \frac{1}{N}\sum_{i=1}^{N}a(i).$$

证明 由定理 1.1，直接可算出

$$E(S) = \sum_{i=1}^{N}c(i)E[a(R_i)] = \bar{a}\sum_{i=1}^{N}c(i) = N\bar{c}\,\bar{a}.$$

注意

$$S - E(S) = \sum_{i=1}^{N}c(i)a(R_i) - N\bar{c}\,\bar{a}$$

$$= \sum_{i=1}^{N}[c(i)-\bar{c}][a(R_i)-\bar{a}],$$

则方差

$$\begin{aligned}
\operatorname{var}(S) &= E\left\{\sum_{i=1}^{N}[c(i)-\bar{c}][a(R_i)-\bar{a}]\right\}^2 \\
&= E\left\{\sum_{i=1}^{N}[c(i)-\bar{c}]^2[a(R_i)-\bar{a}]^2+\sum_{i\neq j}\sum[c(i)-\bar{c}]\right. \\
&\quad\left.\cdot[c(j)-\bar{c}][a(R_i)-\bar{a}][a(R_j)-\bar{a}]\right\} \\
&= \sum_{i=1}^{N}[c(i)-\bar{c}]^2\operatorname{var}[a(R_i)]+\sum_{i\neq j}\sum[c(i)-\bar{c}] \\
&\quad\cdot[c(j)-\bar{c}]\operatorname{cov}[a(R_i),a(R_j)] \\
&= \frac{1}{N}\left\{\sum_{i=1}^{N}[c(i)-\bar{c}]^2\right\}\left\{\sum_{i=1}^{N}[a(i)-\bar{a}]^2\right\}-\frac{1}{N(N-1)} \\
&\quad\cdot\left\{\sum_{i\neq j}\sum[c(i)-\bar{c}][c(j)-\bar{c}]\right\}\left\{\sum_{i=1}^{N}[a(i)-\bar{a}]^2\right\} \\
&= \frac{1}{N}\left\{\sum_{i=1}^{N}[c(i)-\bar{c}]^2\right\}\left\{\sum_{i=1}^{N}[a(i)-\bar{a}]^2\right\} \\
&\quad+\frac{1}{N(N-1)}\left\{\sum_{i=1}^{N}[c(i)-\bar{c}]^2\right\}\left\{\sum_{i=1}^{N}[a(i)-\bar{a}]^2\right\} \\
&= \frac{1}{N-1}\left\{\sum_{i=1}^{N}[c(i)-\bar{c}]^2\right\}\left\{\sum_{i=1}^{N}[a(i)-\bar{a}]^2\right\}.
\end{aligned}$$

证毕.

§3.2 线性秩统计量分布的一些性质

本节我们证明：线性秩统计量对应的分值 $a(1),\cdots,a(N)$ 及回归常数 $c(1),\cdots,c(N)$ 重新排列次序不影响线性秩统计量的分布. 首先引入一个简单的引理.

引理 2.1 若 $K(r)$($r\in\mathscr{R}$, $\mathscr{R}$ 的定义见 (3.1.6) 式)是 $\mathscr{R}$ 到 $\mathscr{R}$ 的一对一映射，R 是在 $\mathscr{R}$ 上均匀分布的秩向量，则统计量 $S=K(R)$ 也在 $\mathscr{R}$ 上均匀分布.

证明 因为$s=K(r)$是$\mathscr{R}$到$\mathscr{R}$的一对一映射，所以存在逆映射$r=K^{-1}(s)$ $(s\in\mathscr{R})$也是$\mathscr{R}$到$\mathscr{R}$的一对一映射. 对任一$s\in\mathscr{R}$,

$$P\{S=s\}=P\{K(R)=s\}=P\{R=K^{-1}(s)\}=\frac{1}{N!}.$$

证毕.

下面我们列举几个常用的一对一映射.

例 1 定义$K_1(r)=d$,其中$d=(d_1,\cdots,d_N)$,$d_i=j$当$r_j=i$ $(i=1,\cdots,N)$时，即d_i是i在$r=(r_1,\cdots,r_N)$中的位置. 例如$N=6$,$r=(4,1,6,2,3,5)$,则

$$K_1(r)=(2,4,5,1,6,3).$$

$K_1(r)$是$\mathscr{R}$到$\mathscr{R}$的一对一映射.

例 2 $K_2(r)=(N+1-r_1,\cdots,N+1-r_N)$也是$\mathscr{R}$到$\mathscr{R}$的一对一映射. 实际上如果样本$X_1,\cdots,X_N$对应的秩向量是$r$,则$-X_1,\cdots,-X_N$对应的秩向量即为$K_2(r)$.

例 3 可以考虑$(1,\cdots,N)$的两个排列

$$r=(r_1,\cdots,r_N) \quad 和 \quad s=(s_1,\cdots,s_N)$$

的复合排列. r对s的复合排列为

$$r\circ s=(r_{s_1},\cdots,r_{s_N}). \tag{3.2.1}$$

而s对r的复合排列为

$$s\circ r=(s_{r_1},\cdots,s_{r_N}). \tag{3.2.2}$$

一般$r\circ s\neq s\circ r$. 例如$r=(2,4,3,1)$,$s=(1,3,4,2)$,则

$$r\circ s=(2,3,1,4),$$
$$s\circ r=(3,2,4,1).$$

对一给定的排列s,映射

$$K_3(r)=r\circ s, \quad r\in\mathscr{R}$$

和

$$K_4(r)=s\circ r, \quad r\in\mathscr{R}$$

都是$\mathscr{R}$到$\mathscr{R}$的一对一映射.

定理 2.2 设 $a(1),\cdots,a(N)$ 和 $c(1),\cdots,c(N)$ 是两组常数. 秩向量 $R=(R_1,\cdots,R_N)$ 在集合 $\mathscr{R}$ 上均匀分布, 则线性秩统计量

$$S=\sum_{i=1}^{N}c(i)a(R_i)$$

与线性秩统计量

$$S'=\sum_{i=1}^{N}a(i)c(R_i)$$

有相同的分布.

证明 由引理 2.1, 由前面例 1 定义的映射 $K_1(r)$ 确定的统计量, $D=K_1(R)$ 在 $\mathscr{R}$ 上均匀分布, 即 D 与 R 同分布. 而

$$S=\sum_{j=1}^{N}c(j)a(R_j)=\sum_{i=1}^{N}c(D_i)a(i)\xlongequal{d}\sum_{i=1}^{N}a(i)c(R_i)=S',$$

其中 $D=(D_1,\cdots,D_N)$. 证毕.

定理 2.3 设 $c'(1),\cdots,c'(N)$ 是 $c(1),\cdots,c(N)$ 的重新排列次序后的一组常数, $a(1),\cdots,a(N)$ 是一组分值常数. 秩向量 $R=(R_1,\cdots,R_N)$ 在集合 $\mathscr{R}$ 上均匀分布, 则线性秩统计量

$$S=\sum_{i=1}^{N}c(i)a(R_i)$$

与线性秩统计量

$$S'=\sum_{i=1}^{N}c'(i)a(R_i)$$

有相同的分布.

证明 设 $(c'(1),\cdots,c'(N))=(c(\alpha_1),\cdots,c(\alpha_N))$, 则统计量

$$S'=\sum_{i=1}^{N}c'(i)a(R_i)=\sum_{i=1}^{N}c(\alpha_i)a(R_i).$$

利用例 1 中的映射 $K_1(r)$, 令 $K_1(\alpha)=d$, 即 $d=(d_1,\cdots,d_N)$, d_j 为 j 在 $\alpha=(\alpha_1,\cdots,\alpha_N)$ 中的位置, 则统计量 S' 可改写为

$$S'=\sum_{i=1}^{N}c(\alpha_i)a(R_i)=\sum_{j=1}^{N}c(j)a(R_{d_j}).$$

而由前面例 3 的映射 $K_3(r)$ 和引理 2.1 知统计量 $Q=R\circ d=$

$(R_{d_1},\cdots,R_{d_N})$在 $\mathscr{R}$ 上均匀分布，即有 $Q \xlongequal{d} R$，所以

$$S' = \sum_{j=1}^{N} c(j)a(R_{d_j}) = \sum_{j=1}^{N} c(j)a(Q_j)$$

$$\xlongequal{d} \sum_{j=1}^{N} c(j)a(R_j) = S.$$

证毕.

由上述定理易于推知：$c'(1),\cdots,c'(N)$是$c(1),\cdots,c(N)$的一个重排，$a'(1),\cdots,a'(N)$是$a(1),\cdots,a(N)$的一个重排. 当秩向量 R 在集合 $\mathscr{R}$ 上均匀分布时，有线性秩统计量

$$S = \sum_{i=1}^{N} c(i)a(R_i) \xlongequal{d} \sum_{i=1}^{N} c'(i)a'(R_i) = S''.$$

定理 2.4 设 $V(R)$是一个秩统计量，若秩向量 $R=(R_1,\cdots,R_N)$在集合 $\mathscr{R}$ 上均匀分布，则 $V(R)$有对称分布的充要条件是：存在 $\mathscr{R}$ 到 $\mathscr{R}$ 的一对一映射 $K(\cdot)$，使对任一 $r\in\mathscr{R}$，有

$$V(r) + V(K(r)) = C, \tag{3.2.3}$$

其中 C 是一常数. $V(R)$的分布的对称点为 $C/2$.

证明 *充分性* 由本节引理 2.1，有

$$K(R) \xlongequal{d} R,$$

则由(3.2.3)式，有

$$V(R) - \frac{C}{2} = \frac{C}{2} - V(K(R)) \xlongequal{d} \frac{C}{2} - V(R).$$

所以 $V(R)$的分布关于 $C/2$ 对称.

必要性 若 $V(R)$关于 μ 对称，对任一使

$$P\{V(R) = \mu + d\} > 0$$

的非负的 d，令集合

$$A_d = \{r \mid r \in \mathscr{R} \text{ 且 } V(r) = \mu + d\},$$

$$B_d = \{r \mid r \in \mathscr{R} \text{ 且 } V(r) = \mu - d\},$$

对不同的 d 与 d'，A_d 与 $A_{d'}$互斥，且对任一 $r\in\mathscr{R}$ 必属于且只属于一个 A_d 或 B_d. 因而不为空集的 A_d, B_d 只有有限个. 记为 $A_{d_1},\cdots,$

A_{d_k}; $B_{d_1},\cdots,B_{d_k}$. 由于秩向量 R 在集合 $\mathscr{R}$ 上均匀分布，统计量 $V(R)$ 又是对称分布的，所以 A_{d_i} 与 B_{d_i} 的元素个数必然一样，因而集合 $\mathscr{R}$ 的有限个元素间可规定一个一对一映射（不一定唯一）如下：

$$\left.\begin{array}{l} K(r)=r' \\ K(r')=r \end{array}\right\} \text{满足 } r \in A_{d_i}, r' \in B_{d_i}, \quad i=1,\cdots,k.$$

对此映射显然有

$$V(r)+V(K(r))=2\mu, \quad \text{对一切 } r \in \mathscr{R}.$$

证毕.

定理 2.5 设

$$S=\sum_{i=1}^{N} c(i) a(R_i)$$

是线性秩统计量. 以 $c'(1)\leqslant\cdots\leqslant c'(N)$ 和 $a'(1)\leqslant\cdots\leqslant a'(N)$ 分别记 $c(1),\cdots,c(N)$ 和 $a(1),\cdots,a(N)$ 按值的大小排列后的值. 若秩向量 $R=(R_1,\cdots,R_N)$ 在集合 $\mathscr{R}$ 上均匀分布，且有

$$a'(i)+a'(N+1-i)=\text{常数}, \quad i=1,\cdots,N$$

或

$$c'(i)+c'(N+1-i)=\text{常数}, \quad i=1,\cdots,N,$$

则线性秩统计量 S 的分布关于 $N\bar{c}\bar{a}$ 对称.

证明 取 $\mathscr{R}$ 到 $\mathscr{R}$ 的一对一映射

$$K(r)=(N+1-r_1,\cdots,N+1-r_N), \quad r=(r_1,\cdots,r_N)\in\mathscr{R}$$

并利用本节定理 2.4 及定理 2.3，定理 2.2 即得本定理结论. 证毕.

例 4 考虑 Wilcoxon 秩和统计量（参见第一章第 3 节）

$$W=\sum_{i=1}^{n} R_i,$$

其对应的两组常数为

$$c(i)=\begin{cases} 0, & \text{当 } i=1,\cdots,m \text{ 时}, \\ 1, & \text{当 } i=m+1,\cdots,N \text{ 时}; \end{cases}$$

$$a(i)=i,\quad i=1,\cdots,N.$$

易见

$$a(i)+a(N+1-i)=N+1,\quad i=1,\cdots,N.$$

故当秩向量 $R=(Q_1,\cdots,Q_m,R_1,\cdots,R_n)$在 $\mathscr{R}$ 上均匀分布时,W 的分布关于

$$N\bar{c}\,\bar{a}=N\cdot\frac{n}{N}\cdot\frac{N+1}{2}=\frac{n(N+1)}{2}$$

对称.

例 5 考虑线性秩统计量

$$S=\sum_{i=1}^{N}iR_i,$$

其对应的

$$c(i)=i,\quad a(i)=i,\quad i=1,\cdots,N.$$

故当秩向量 $R=(R_1,\cdots,R_N)$在 $\mathscr{R}$ 上均匀分布时,S 的分布关于 $N\left(\frac{N+1}{2}\right)^2$对称.

§3.3 讨论线性秩统计量渐近性质的一些预备定理

本节和下一节讨论线性秩统计量在 H_0 假设下,即秩向量 $R=(R_1,\cdots,R_N)$在集合 $\mathscr{R}$ 上均匀分布时,线性秩统计量的渐近正态性. 由于定理较多,故分为两节. 为了讨论渐近性质,记线性秩统计量

$$S_N=\sum_{i=1}^{N}c_N(i)a_N(R_i),\tag{3.3.1}$$

对分值和回归常数 $a_N(i)$和 $c_N(i)$加上附标 N,表示它们是样本量 N 时的两组数. 在下面的讨论中,对这两组数作一些限制. 对回归常数$\{c_N(i)\}$要求满足 Noether 条件,即要求当 $N\to\infty$时,有

$$\frac{\sum_{i=1}^{N}[c_N(i)-\bar{c}_N]^2}{\max\limits_{1\leqslant i\leqslant N}[c_N(i)-\bar{c}_N]^2}\to\infty, \tag{3.3.2}$$

其中$\bar{c}_N=\frac{1}{N}\sum_{i=1}^{N}c_N(i)$. 这就是说$\{c_N(i)\}$没有特别显著的离群值.

对分值$\{a_N(i)\}$要求

$$a_N(i)=b_N\phi\left(\frac{i}{N+1}\right)+d_N,\quad i=1,\cdots,N, \tag{3.3.3}$$

其中 b_N,d_N 是仅依赖 N,而不随 i 变化的常数. 而 $\phi\left(\frac{i}{N+1}\right)$ 是函数 $\phi(u)$在$\frac{i}{N+1}$ $(i=1,\cdots,N)$处的取值. 函数 $\phi(u)$满足:

(i) $\phi(u)$是定义在$(0,1)$上的函数,不随 N 变化;

(ii) $\phi(u)$可表成两个非降函数之差,即

$$\phi(u)=\phi'(u)-\phi''(u).$$

而 $\phi'(u)$与 $\phi''(u)$均是 u 的非降函数.

(iii) $\phi(u)$平方可积,即

$$0<\int_0^1[\phi(u)-\bar{\phi}]^2\mathrm{d}u<\infty,\quad 其中\ \bar{\phi}=\int_0^1\phi(u)\mathrm{d}u.$$

今后称满足(i),(ii),(iii)的函数 $\phi(u)$为平方可积分值函数.

由于线性秩统计量

$$\begin{aligned}S_N'&=\sum_{i=1}^{N}c_N(i)\left[b_N\phi\left(\frac{R_i}{N+1}\right)+d_N\right]\\&=b_N\sum_{i=1}^{N}c_N(i)\phi\left(\frac{R_i}{N+1}\right)+d_NN\bar{c}_N\\&\xlongequal{\text{def}}b_NS_N+d_NN\bar{c}_N,\end{aligned}$$

从而 S_N' 和 S_N 两个线性秩统计量间有下列等式:

$$\frac{S_N'-E(S_N')}{\sqrt{\mathrm{var}S_N'}}=\frac{S_N-E(S_N)}{\sqrt{\mathrm{var}S_N}},$$

因而讨论渐近性质时,我们不妨假定

$$b_N = 1,\quad d_N = 0,\quad a_N(i) = \phi\left(\frac{i}{N+1}\right).$$

例 1 考虑回归常数

$$c_N(i) = \begin{cases} 0, & i = 1,\cdots,m, \\ 1, & i = m+1,\cdots,N, N = m+n. \end{cases} \tag{3.3.4}$$

此时，$\bar{c}_N = n/N$，

$$\sum_{i=1}^{N}[c_N(i) - \bar{c}_N]^2 = \frac{mn^2}{N^2} + \frac{nm^2}{N^2} = \frac{mn}{N},$$

$$\max_{1\leqslant i\leqslant N}[c_N(i) - \bar{c}_N]^2 = \frac{[\max(m,n)]^2}{N^2},$$

则(3.3.2)式的 Noether 条件为

$$\frac{\sum_{i=1}^{N}[c_N(i) - \bar{c}_N]^2}{\max_{1\leqslant i\leqslant N}[c(i) - \bar{c}_N]^2} = \frac{Nnm}{[\max(m,n)]^2} = \frac{N\cdot\min(m,n)}{\max(m,n)}$$

$$= \left[1 + \frac{\min(m,n)}{\max(m,n)}\right]\cdot\min(m,n) \to \infty.$$

所以条件相当于 $\min(m,n)\to\infty$.

又考虑回归常数

$$c_N(i) = i \quad (i = 1,\cdots,N), \tag{3.3.5}$$

则 $\bar{c}_N = \dfrac{N+1}{2}$，

$$\sum_{i=1}^{N}[c_N(i) - \bar{c}_N]^2 = \sum_{i=1}^{N}\left[i - \frac{N+1}{2}\right]^2 = \frac{N(N-1)(N+1)}{12},$$

$$\max_{1\leqslant i\leqslant N}[c_N(i) - \bar{c}_N]^2 = \frac{(N-1)^2}{4}.$$

此时(3.3.2)式的 Noether 条件自然成立：

$$\frac{\sum_{i=1}^{N}[c_N(i) - \bar{c}_N]^2}{\max_{1\leqslant i\leqslant N}[c_N(i) - \bar{c}_N]^2} = \frac{N(N-1)(N+1)}{3(N-1)^2} \to \infty.$$

例 2 考虑分值 $a(i)=i$ $(i=1,\cdots,N)$，则

$$a(i) = (N+1)\cdot\left(\frac{i}{N+1}\right), \quad i = 1,\cdots,N.$$

对应的平方可积分值函数为 $\phi(u)=u$, $b_N=(N+1)$.

又考虑分值 $a_N(i)=\left(i-\frac{N+1}{2}\right)^2=(N+1)^2\left(\frac{i}{N+1}-\frac{1}{2}\right)^2$ $(i=1,\cdots,N)$,则对应的 $b_N=(N+1)^2$,分值函数为

$$\phi(u) = \left(u-\frac{1}{2}\right)^2,$$

而 $\phi(u)=\phi'(u)-\phi''(u)$,这里

$$\phi'(u) = \begin{cases} 0, & 0<u<\frac{1}{2}, \\ \left(u-\frac{1}{2}\right)^2, & \frac{1}{2}\leqslant u<1; \end{cases}$$

$$\phi''(u) = \begin{cases} -\left(u-\frac{1}{2}\right)^2, & 0<u<\frac{1}{2}, \\ 0, & \frac{1}{2}\leqslant u<1 \end{cases}$$

都是(0,1)上的非降函数. 以上的 $\phi(u)$ 均满足条件(i)~(iii),是平方可积分值函数.

引理 3.1 设$\{X_n\}$,$\{Y_n\}$是两列随机变量,若有

$$\lim_{n\to\infty}E(X_n-Y_n)^2=0,$$

则

$$\lim_{n\to\infty}E(X_n)=\lim_{n\to\infty}E(Y_n), \quad \lim_{n\to\infty}E(X_n^2)=\lim_{n\to\infty}E(Y_n^2).$$

上述等式当式中的量存在时成立.

证明 对随机变量 X_n, Y_n 有

$$[E(X_n)-E(Y_n)]^2=[E(X_n-Y_n)]^2\leqslant E(X_n-Y_n)^2,$$

所以

$$\lim_{n\to\infty}[E(X_n)-E(Y_n)]=0.$$

故当 $\lim_{n\to\infty}E(X_n)$ 或 $\lim_{n\to\infty}E(Y_n)$ 存在时,有

$$\lim_{n\to\infty}E(X_n)=\lim_{n\to\infty}E(Y_n).$$

又由于

$$\begin{aligned}\operatorname{var}(X_n - Y_n) &= \operatorname{var}X_n + \operatorname{var}Y_n - 2\operatorname{cov}(X_n, Y_n) \\ &\geqslant \operatorname{var}X_n + \operatorname{var}Y_n - 2\sqrt{\operatorname{var}X_n \cdot \operatorname{var}Y_n} \\ &= \left[\sqrt{\operatorname{var}X_n} - \sqrt{\operatorname{var}Y_n}\right]^2,\end{aligned}$$

而

$$\operatorname{var}(X_n - Y_n) \leqslant E(X_n - Y_n)^2,$$

所以当 $\lim\limits_{n\to\infty} E(X_n - Y_n)^2 = 0$ 时,有

$$\lim_{n\to\infty}\left[\sqrt{\operatorname{var}X_n} - \sqrt{\operatorname{var}Y_n}\right] = 0.$$

当 $\lim\limits_{n\to\infty}\operatorname{var}X_n$ 或 $\lim\limits_{n\to\infty}\operatorname{var}Y_n$ 存在时,则有

$$\lim_{n\to\infty}\operatorname{var}X_n = \lim_{n\to\infty}\operatorname{var}Y_n,$$

从而有

$$\lim_{n\to\infty} E(X_n^2) = \lim_{n\to\infty} E(Y_n^2).$$

证毕.

引理 3.2 设 $\phi(u)$是平方可积分值函数,

$$a_N(i) = \phi\left(\frac{i}{N+1}\right), \quad i = 1, \cdots, N.$$

若秩统计量 R_1 有

$$P\{R_1 = i\} = \frac{1}{N}, \quad i = 1, \cdots, N,$$

则

$$\lim_{N\to\infty} E\left[\phi\left(\frac{R_1}{N+1}\right)\right]^2 = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} a_N^2(i) = \int_0^1 \phi^2(u)\mathrm{d}u. \tag{3.3.6}$$

证明 取函数

$$g_N(u) = \begin{cases}\phi\left(\dfrac{i}{N+1}\right), & \dfrac{i}{N+1} \leqslant u \leqslant \dfrac{i+1}{N+1}, i = 1, \cdots, N, \\ \phi(u), & 0 < u < \dfrac{1}{N+1},\end{cases}$$

则

$$\int_0^1 g_N^2(u)\mathrm{d}u = \int_0^{\frac{1}{N+1}} \phi^2(u)\mathrm{d}u + \frac{N}{N+1}\left[\frac{1}{n}\sum_{i=1}^{N} a_N^2(i)\right].$$

因为 $\phi(u)$是平方可积的，所以有

$$\lim_{N\to\infty}\int_0^{\frac{1}{N+1}} \phi^2(u)\mathrm{d}u = 0.$$

故若能证明

$$\lim_{N\to\infty}\int_0^1 g_N^2(u)\mathrm{d}u = \int_0^1 \phi^2(u)\mathrm{d}u,$$

则有引理结论. 而由本节引理 3.1，只要能证明下式即可：

$$\lim_{N\to\infty}\int_0^1 [g_N(u) - \phi(u)]^2\mathrm{d}u = 0.$$

下面我们来证明上述关系式. 首先分值函数

$\phi(u) = \phi'(u) - \phi''(u) = \phi'_+(u) - \phi'_-(u) - \phi''_+(u) + \phi''_-(u)$，其中 $\phi'(u)$，$\phi''(u)$是两个非降函数. $\phi'_+(u)$，$\phi'_-(u)$，$\phi''_+(u)$，$\phi''_-(u)$是它们的正部和负部，$\phi'_+(u)$与 $\phi''_+(u)$是非负非降函数，$\phi'_-(u)$与 $\phi''_-(u)$是非负非升函数. 而 $g_N(u)$亦可对应地分解为

$$g_N(u) = g'_{N+}(u) - g'_{N-}(u) - g''_{N+}(u) + g''_{N-}(u).$$

又

$$\begin{aligned}
&[g_N(u) - \phi(u)]^2 \\
&\quad\leqslant 2[g'_N(u) - \phi'(u)]^2 + 2[g''_N(u) - \phi''(u)]^2 \\
&\quad\leqslant 4[g'_{N+}(u) - \phi'_+(u)]^2 + 4[g'_{N-}(u) - \phi'_-(u)]^2 \\
&\qquad + 4[g''_{N+}(u) - \phi''_+(u)]^2 + 4[g''_{N-}(u) - \phi''_-(u)]^2,
\end{aligned}$$

故我们只需对 $\phi(u)$是非负单调函数证明关系式

$$\lim_{N\to\infty}\int_0^1 [g_N(u) - \phi(u)]^2\mathrm{d}u = 0.$$

若 $\phi(u)$在(0，1)上非负单调可积，它至多有可数个跳跃点，从而除可数个点外，有

$$\lim_{N\to\infty} g_N(u) = \phi(u), \quad \lim_{N\to\infty} g_N^2(u) = \phi^2(u),$$

$$\lim_{N\to\infty} g_N(u)\phi(u) = \phi^2(u).$$

当 $\phi(u)$ 非负非降时，则有 $g_N(u) \leqslant \phi(u)$，由控制收敛定理即可得

$$\lim_{N\to\infty}\int_0^1 g_N^2(u)\mathrm{d}u = \int_0^1 \phi^2(u)\mathrm{d}u,$$

$$\lim_{N\to\infty}\int_0^1 g_N(u)\phi(u)\mathrm{d}u = \int_0^1 \phi^2(u)\mathrm{d}u,$$

从而

$$\lim_{N\to\infty}\int_0^1 [g_N(u) - \phi(u)]^2\mathrm{d}u = 0.$$

当 $\phi(u)$ 非负非升时，有

$$\int_0^1 g_N^2(u)\mathrm{d}u \leqslant \int_0^{\frac{1}{N+1}} \phi^2(u)\mathrm{d}u + \int_0^1 \phi^2(u)\mathrm{d}u - \int_{\frac{N}{N+1}}^1 \phi^2(u)\mathrm{d}u,$$

两边取上极限即得

$$\overline{\lim_{N\to\infty}}\int_0^1 g_N^2(u)\mathrm{d}u \leqslant \int_0^1 \phi^2(u)\mathrm{d}u.$$

又由于 $\phi(u)$ 非升，$g_N(u) \geqslant \phi(u)$，$u \in (0,1)$. 根据 Fatou 引理，有

$$\int_0^1 \phi^2(u)\mathrm{d}u \leqslant \underline{\lim_{N\to\infty}}\int_0^1 g_N^2(u)\mathrm{d}u.$$

所以综合上面两式即得

$$\lim_{N\to\infty}\int_0^1 g_N^2(u)\mathrm{d}u = \int_0^1 \phi^2(u)\mathrm{d}u.$$

类似地可证

$$\lim_{N\to\infty}\int_0^1 g_N(u)\phi(u)\mathrm{d}u = \int_0^1 \phi^2(u)\mathrm{d}u,$$

从而亦有

$$\lim_{N\to\infty}\int_0^1 [g_N(u) - \phi(u)]^2\mathrm{d}u = 0.$$

综合上述论证即知引理结论成立. 证毕.

引理 3.3 对二项分布

$$f_N(i)=\binom{N}{i}p^i(1-p)^{N-i},\quad i=0,1,\cdots,N$$

有：当 $i>(N+1)p-1$ 时，$f_N(i)>f_N(i+1)$；当 $i<(N+1)p$ 时，$f_N(i-1)<f_N(i)$，即二项分布的众数 m_0 满足

$$p(N+1)-1\leqslant m_0\leqslant p(N+1).$$

证明 考虑比式：

$$\frac{f_N(i+1)}{f_N(i)}=\frac{\binom{N}{i+1}p^{(i+1)}(1-p)^{N-(i+1)}}{\binom{N}{i}p^i(1-p)^{N-i}}=\frac{(N-i)p}{(i+1)(1-p)},$$

当 $i>(N+1)p-1$ 时，$\dfrac{(N-i)p}{(i+1)(1-p)}<1$，故 $f_N(i)>f_N(i+1)$；当 $i<(N+1)p$ 时，$\dfrac{[N-(i-1)]p}{ip}>1$，故 $f_N(i-1)<f_N(i)$。证毕.

引理 3.4 设 $b(1)\geqslant\cdots\geqslant b(K)$ 和 $a(1)\leqslant\cdots\leqslant a(K)$ 是两组常数，则

$$\sum_{i=1}^{K}b(i)a(i)\leqslant K\bar{b}\cdot\bar{a},\tag{3.3.7}$$

其中 $\bar{b}=\dfrac{1}{K}\sum\limits_{i=1}^{K}b(i)$，$\bar{a}=\dfrac{1}{K}\sum\limits_{i=1}^{K}a(i)$.

证明 (3.3.7)式移项后，可写成

$$\sum_{i=1}^{K}[b(i)-\bar{b}]a(i)\leqslant 0.$$

设 $b(j)\geqslant\bar{b}\geqslant b(j+1)$，则当 $i\leqslant j$ 时，$b(i)-\bar{b}\geqslant 0$；$i\geqslant j+1$ 时，$b(i)-\bar{b}\leqslant 0$. 记 a' 是一常数，满足 $a(j)\leqslant a'\leqslant a(j+1)$，则

$$\begin{aligned}\sum_{i=1}^{K}[b(i)-\bar{b}]a(i)&=\sum_{i=1}^{j}[b(i)-\bar{b}]a(i)+\sum_{i=j+1}^{K}[b(i)-\bar{b}]a(i)\\&\leqslant\sum_{i=1}^{j}[b(i)-\bar{b}]a'+\sum_{i=j+1}^{K}[b(i)-\bar{b}]a'\\&=a'\sum_{i=1}^{K}[b(i)-\bar{b}]=0.\end{aligned}$$

故引理结论成立. 证毕.

引理 3.5 设 $\phi(u)$是平方可积分值函数，Y_N 是参数为 N、p 的二项分布的随机变量，则

$$\lim_{N\to\infty} E\left[\phi\left(\frac{Y_N+1}{N+2}\right)\right]=\phi(p) \qquad (3.3.8)$$

对 $\phi(u)$的连续点 $p\in(0,1)$成立。

证明 因为 $\phi(u)$是平方可积分值函数，它可表为两个非降函数的差，故不失一般性可假定 $\phi(u)$非降来证明引理. $\phi(u)$在$(0,1)$上非降，至多有可数个不连续点. 对任一 $\phi(u)$的连续点 $p\in(0,1)$，对任给 $\varepsilon>0$，存在 $d_\varepsilon>0$，使 $|u-p|\leqslant d_\varepsilon$ 时，有

$$|\phi(u)-\phi(p)|\leqslant\varepsilon.$$

以 $f_N(i)$ $(i=0,1,\cdots,N)$和 $F_N(x)$ $(-\infty<x<\infty)$分别记参数为 N、p 的二项分布的概率函数和分布函数，则

$$\begin{aligned}
&E\left[\phi\left(\frac{Y_N+1}{N+2}\right)-\phi(p)\right]^2\\
&\quad=\int_{\{y=0,\cdots,N\}}\left[\phi\left(\frac{y+1}{N+2}\right)-\phi(p)\right]^2\mathrm{d}F_N(y)\\
&\quad=\int_{\left\{\left|\frac{y+1}{N+2}-p\right|\leqslant d_\varepsilon,y=0,\cdots,N\right\}}\left[\phi\left(\frac{y+1}{N+2}\right)-\phi(p)\right]^2\mathrm{d}F_N(y)\\
&\qquad+\int_{\left\{\left|\frac{y+1}{N+2}-p\right|>d_\varepsilon,y=0,\cdots,N\right\}}\left[\phi\left(\frac{y+1}{N+2}\right)-\phi(p)\right]^2\mathrm{d}F_N(y)\\
&\quad\leqslant\varepsilon^2+\int_{\left\{\left|\frac{y+1}{N+2}-p\right|>d_\varepsilon,y=0,\cdots,N\right\}}\left[\phi\left(\frac{y+1}{N+2}\right)-\phi(p)\right]^2\mathrm{d}F_N(y).
\end{aligned}$$

考虑最后这个积分的一部分

$$\begin{aligned}
&\int_{\left\{\left(\frac{y+1}{N+2}-p\right)>d_\varepsilon,y=0,\cdots,N\right\}}\left[\phi\left(\frac{y+1}{N+2}\right)-\phi(p)\right]^2\mathrm{d}F_N(y)\\
&\quad=\int_{\{y>(N+2)(p+d_\varepsilon)-1,y=0,\cdots,N\}}\left[\phi\left(\frac{y+1}{N+2}\right)-\phi(p)\right]^2\mathrm{d}F_N(y)\\
&\quad=\sum_{i=K_N}^{N}\left[\phi\left(\frac{i+1}{N+2}\right)-\phi(p)\right]^2 f_N(i),
\end{aligned}$$

其中 $K_N=[(N+2)(p+d_\epsilon)]$ 为不大于 $(N+2)(p+d_\epsilon)$ 的最大整数. 当 $i\geqslant K_N$ 时, $f_N(i)$ 随 i 上升而下降, 而 $\left[\phi\left(\frac{i+1}{N+2}\right)-\phi(p)\right]^2$ 随 i 上升而上升. 故由本节引理 3.4, 有

$$\sum_{i=K_N}^{N}\left[\phi\left(\frac{i+1}{N+2}\right)-\phi(p)\right]^2 f_N(i)$$

$$\leqslant \bar{f}_N(K_N,N)\cdot\sum_{i=K_N}^{N}\left[\phi\left(\frac{i+1}{N+2}\right)-\phi(p)\right]^2$$

$$\leqslant (N+2)\,\bar{f}_N(K_N,N)\int_{\frac{K_N+1}{N+2}}^{1}[\phi(u)-\phi(p)]^2\mathrm{d}u$$

$$\leqslant (N+2)\,\bar{f}_N(K_N,N)\int_{0}^{1}[\phi(u)-\phi(p)]^2\mathrm{d}u,$$

其中

$$\bar{f}_N(K_N,N)=\frac{1}{N-K_N+1}\sum_{i=K_N}^{N}f_N(i).$$

而

$$(N+2)\,\bar{f}_N(K_N,N)=\frac{N+2}{N-K_N+1}P\left\{\frac{Y_N}{N}>\frac{K_N-1}{N}\right\},$$

但

$$\frac{K_N}{N}\to p+d_\epsilon,\quad \frac{Y_N}{N}\xrightarrow{P} p,$$

所以当 $N\to\infty$ 时,

$$P\left\{\frac{Y_N}{N}>\frac{K_N-1}{N}\right\}\to 0.$$

另外可取 $d_\epsilon<1-p$, 那么

$$\frac{N+2}{N-K_N+1}\to\frac{1}{1-(p+d_\epsilon)}>0.$$

结合上面两式可得

$$\lim_{N\to\infty}(N+2)\,\bar{f}_N(K_N,N)=0,$$

从而证得: 当 $N\to\infty$ 时,

$$\int_{\left\{\left(\frac{y+1}{N+2}-p\right)>d_\varepsilon, y=0,\cdots,N\right\}}\left[\phi\left(\frac{y+1}{N+2}\right)-\phi(p)\right]^2 \mathrm{d}F_N(y)\to 0.$$

类似地可证：当 $N\to\infty$ 时，

$$\int_{\left\{\left(\frac{y+1}{N+2}-p\right)<-d_\varepsilon, y=0,\cdots,N\right\}}\left[\phi\left(\frac{y+1}{N+2}\right)-\phi(p)\right]^2 \mathrm{d}F_N(y)\to 0.$$

因而对任给 $\delta>0$，当 N 足够大时，可使

$$E\left[\phi\left(\frac{Y_N+1}{N+2}\right)-\phi(p)\right]^2\leqslant\delta,$$

即

$$\lim_{N\to\infty}E\left[\phi\left(\frac{Y_N+1}{N+2}\right)-\phi(p)\right]^2=0.$$

由本节引理 3.1 即得

$$\lim_{N\to\infty}E\left[\phi\left(\frac{Y_N+1}{N+2}\right)\right]=\phi(p).$$

证毕.

定理 3.6 若 $\phi(u)$ 是平方可积分值函数，秩向量 $R=(R_1,\cdots,R_N)$ 是 $(0,1)$ 均匀分布样本 $U_1,\cdots,U_N$ 的秩向量，则

$$\lim_{N\to\infty}E\left[\phi\left(\frac{R_1}{N+1}\right)-\phi(U_1)\right]^2=0. \qquad (3.3.9)$$

证明 不失一般性可设 $\phi(u)$ 非负.

$$E\left[\phi\left(\frac{R_1}{N+1}\right)-\phi(U_1)\right]^2$$

$$=E\phi^2\left(\frac{R_1}{N+1}\right)+E\phi^2(U_1)-2E\phi\left(\frac{R_1}{N+1}\right)\phi(U_1),$$

其中

$$E\phi^2(U_1)=\int_0^1\phi^2(u)\mathrm{d}u.$$

由本节引理 3.2 有

$$\lim_{N\to\infty}E\phi^2\left(\frac{R_1}{N+1}\right)=\int_0^1\phi^2(u)\mathrm{d}u,$$

故只需证明

$$\lim_{N\to\infty} E\phi\left(\frac{R_1}{N+1}\right)\phi(U_1) = \int_0^1 \phi^2(u)\mathrm{d}u.$$

下面证明上述关系式. 由于有秩统计量 $R_1 = 1 + \sum_{i=2}^{N}\psi(U_1 - U_i)$，其中函数 $\psi(\cdot)$ 的定义参见(1.2.1)式. 记 $Y(u, N-1) = \sum_{i=2}^{N}\psi(u - U_i)$，它的分布是参数 $N-1$ 与 u 的二项分布，且它与 U_1 独立. 故

$$\begin{aligned} b_N(u) &\xlongequal{\text{def}} E\left[\phi\left(\frac{R_1}{N+1}\right)\Big|U_1 = u\right] \\ &= E\left[\phi\left(\frac{1 + \sum_{i=2}^{N}\psi(u - U_i)}{N+1}\right)\Big|U_1 = u\right] \\ &= E\left[\phi\left(\frac{1 + Y(u, N-1)}{N+1}\right)\right], \end{aligned}$$

则

$$\begin{aligned} E\left[\phi\left(\frac{R_1}{N+1}\right)\phi(U_1)\right] &= E_{U_1}\left\{E\left[\phi\left(\frac{R_1}{N+1}\right)\phi(U_1)\Big|U_1\right]\right\} \\ &= \int_0^1 b_N(u)\phi(u)\mathrm{d}u. \end{aligned}$$

由本节引理 3.5，除可数点外，在(0,1)上，有

$$\left.\begin{aligned} &b_N(u) \to \phi(u), \\ &b_N(u)\phi(u) \to \phi^2(u), \end{aligned}\right. \quad N \to \infty.$$

由假设 $\phi(u)$ 非负，有 $b_N(u)\phi(u) \geqslant 0$，则由 Fatou 引理，有

$$\varliminf_{N\to\infty}\int_0^1 b_N(u)\phi(u)\mathrm{d}u \geqslant \int_0^1 \phi^2(u)\mathrm{d}u.$$

另一方面，

$$\begin{aligned} \int_0^1 b_N^2(u)\mathrm{d}u &= \int_0^1 \left\{E\left[\phi\left(\frac{1 + Y(u, N-1)}{N+1}\right)\right]\right\}^2 \mathrm{d}u \\ &\leqslant \int_0^1 E\left[\phi^2\left(\frac{1 + Y(u, N-1)}{N+1}\right)\right]\mathrm{d}u \\ &= E_{U_1}E\left[\phi^2\left(\frac{R_1}{N+1}\right)\Big|U_1\right] \end{aligned}$$

$$= E\phi^2\left(\frac{R_1}{N+1}\right),$$

故

$$\overline{\lim_{N\to\infty}}\int_0^1 b_N^2(u)\mathrm{d}u \leqslant \overline{\lim_{N\to\infty}} E\left[\phi^2\left(\frac{R_1}{N+1}\right)\right] = \int_0^1 \phi^2(u)\mathrm{d}u,$$

$$\overline{\lim_{N\to\infty}}\int_0^1 b_N(u)\phi(u)\mathrm{d}u \leqslant \overline{\lim_{N\to\infty}}\ \frac{1}{2}\left[\int_0^1 b_N^2(u)\mathrm{d}u + \int_0^1 \phi^2(u)\mathrm{d}u\right]$$

$$\leqslant \int_0^1 \phi^2(u)\mathrm{d}u.$$

从而综合上列各式,即得

$$\lim_{N\to\infty} E\left[\phi\left(\frac{R_1}{N+1}\right)\phi(U_1)\right] = \int_0^1 \phi^2(u)\mathrm{d}u.$$

定理得证.证毕.

§3.4 线性秩统计量在 H_0 下的渐近正态性

本节要证明:线性秩统计量

$$S_N = \sum_{i=1}^N c_N(i)a_N(R_i) = \sum_{i=1}^N [c_N(i) - \bar{c}_N]a_N(R_i) + N\bar{c}_N\cdot\bar{a}_N$$

$$= \sum_{i=1}^N [c_N(i) - \bar{c}_N]\phi\left(\frac{R_i}{N+1}\right) + N\bar{c}_N\cdot\bar{a}_N,$$

其中 $a_N(i)=\phi\left(\frac{i}{N+1}\right)$,$\bar{c}_N=\frac{1}{N}\sum_{i=1}^N c_N(i)$,$\bar{a}_N=\frac{1}{N}\sum_{i=1}^N a_N(i)$,当秩向量 $R=(R_1,\cdots,R_N)$ 在集合 $\mathscr{R}$ 上均匀分布时,在适当的条件下,S_N 与统计量

$$V_N = \sum_{i=1}^N [c_N(i) - \bar{c}_N]\phi(U_i) + N\bar{c}_N\bar{a}_N \qquad (3.4.1)$$

有相同的渐近分布,而 V_N 正则化后的渐近分布为 $N(0,1)$.

定理 4.1 定义统计量

$$V_N = \sum_{i=1}^N [c_N(i) - \bar{c}_N]\phi(U_i) + N\bar{c}_N\bar{a}_N,$$

其中$\{c_N(i)\}$满足线性秩统计量回归常数的 Noether 条件,$\phi(u)$为平方可积分值函数,$U_1,\cdots,U_N$ 是独立同分布(0,1)均匀分布的随机变量列,则当 $N\to\infty$时,

$$\frac{V_N - E(V_N)}{\sqrt{\mathrm{var}V_N}} \text{ 有渐近正态分布 } N(0,1).$$

证明　$\{V_N\}$是独立随机变量和的序列.

$$E(V_N) = \sum_{i=1}^{N}[c_N(i) - \bar{c}_N]\int_0^1 \phi(u)\mathrm{d}u + N\bar{c}_N\bar{a}_N$$
$$= N\bar{c}_N\bar{a}_N, \tag{3.4.2}$$

$$\mathrm{var}V_N = \sum_{i=1}^{N}[c_N(i) - \bar{c}_N]^2 \mathrm{var}\phi(U_i)$$
$$= \sum_{i=1}^{N}[c_N(i) - \bar{c}_N]^2 \cdot \int_0^1[\phi(u) - \bar{\phi}]^2\mathrm{d}u, \tag{3.4.3}$$

其中$\bar{\phi}=\int_0^1\phi(u)\mathrm{d}u$. 而独立随机变量序列$\{[c_N(i)-\bar{c}_N]\phi(U_i)\}$满足著名的 Lindeberg 中心极限定理的条件,即对任给 $\epsilon>0$,有

$$\lim_{N\to\infty}\frac{1}{\mathrm{var}V_N}\sum_{i=1}^{N}\int_{\{u \mid |c_N(i)-\bar{c}_N|\cdot|\phi(u)-\bar{\phi}|>\epsilon\sqrt{\mathrm{var}V_N}\}}[c_N(i) - \bar{c}_N]^2 \cdot [\phi(u) - \bar{\phi}]^2\mathrm{d}u = 0,$$

这是因为

$$\frac{1}{\mathrm{var}V_N}\sum_{i=1}^{N}\int_{\{u \mid |c_N(i)-\bar{c}_N|\cdot|\phi(u)-\bar{\phi}|>\epsilon\sqrt{\mathrm{var}V_N}\}}[c_N(i) - \bar{c}_N]^2 \cdot [\phi(u) - \bar{\phi}]^2\mathrm{d}u$$
$$\leqslant \frac{1}{\mathrm{var}V_N}\sum_{i=1}^{N}[c_N(i) - \bar{c}_N]^2 \cdot \int_{\{u \mid |\phi(u)-\bar{\phi}|>\epsilon\delta_N\}}[\phi(u) - \bar{\phi}]^2\mathrm{d}u$$
$$= \left\{\int_0^1[\phi(u) - \bar{\phi}]^2\mathrm{d}u\right\}^{-1} \cdot \int_{\{u \mid |\phi(u)-\bar{\phi}|>\epsilon\delta_N\}}[\phi(u) - \bar{\phi}]^2\mathrm{d}u,$$

式中的

$$\delta_N = \frac{\sqrt{\mathrm{var}V_N}}{\max\limits_{1\leqslant i\leqslant N}|c_N(i) - \bar{c}_N|}$$

$$= \left[\int_0^1[\phi(u) - \bar{\phi}]^2\mathrm{d}u \cdot \frac{\sum\limits_{i=1}^{N}[c_N(i) - \bar{c}_N]^2}{\max[c_N(i) - \bar{c}_N]^2}\right]^{\frac{1}{2}}$$

$$\to \infty, \quad \text{当 } N \to \infty \text{ (Noether 条件)},$$

故 N 足够大时,集合 $\{u \mid \phi(u) - \bar{\phi}| > \epsilon\delta_N\}$ 的勒贝格测度可任意小,而 $\phi(u)$ 平方可积,所以前述 Lindeberg 中心极限定理的条件满足,从而 $(V_N - E(V_N))/\sqrt{\mathrm{var}V_N}$ 有渐近正态分布 $N(0,1)$.

定理 4.2 线性秩统计量

$$S_N = \sum_{i=1}^{N} c_N(i)a_N(R_i),$$

其中回归常数 $\{c_N(i)\}$ 满足 Noether 条件,分值 $a_N(i) = \phi\left(\frac{i}{N+1}\right)$, $\phi(u)$ 是平方可积分值函数. 若秩向量 $R = (R_1, \cdots, R_N)$ 在集合 $\mathscr{R}$ 上均匀分布,则

$$\frac{S_N - \mu_N}{\sigma_N} \text{ 有渐近正态分布 } N(0,1),$$

其中

$$\mu_N = E(S_N) = N\bar{c}_N\bar{a}_N, \tag{3.4.4}$$

$$\sigma_N^2 = \mathrm{var}S_N = \frac{1}{N-1}\left\{\sum_{i=1}^{N}[c_N(i) - \bar{c}_N]^2\right\}\left\{\sum_{i=1}^{N}[a_N(i) - \bar{a}_N]^2\right\}. \tag{3.4.5}$$

证明 可假定秩向量 R 是 $(0,1)$ 均匀分布的样本 $U_1, \cdots, U_N$ 的秩向量. 由第一章第 6 节定理 6.2 知 R 与次序统计量 $U^0 = (U_{(1)}, \cdots, U_{(N)})$ 独立,从而给定 $U^0 = u$ 时,R 仍在 $\mathscr{R}$ 上均匀分布,记 $u = (u_1, \cdots, u_N)$,则

$$E[(S_N - V_N)^2 | U^0 = u]$$

$$= E\left[\left\{\sum_{i=1}^{N}[c_N(i) - \bar{c}_N][a_N(R_i) - \phi(U_i)]\right\}^2 \Big| U^0 = u\right]$$

$$= E\left[\left\{\sum_{i=1}^{N}[c_N(i) - \bar{c}_N][a_N(R_i) - \phi(u_{(R_i)})]\right\}^2\right]$$

$$= \frac{1}{N-1}\left\{\sum_{i=1}^{N}[c_N(i) - \bar{c}_N]^2\right\}$$

$$\cdot \left\{\sum_{i=1}^{N}[a_N(i) - \phi(u_{(i)}) - \bar{a}_N + \bar{\phi}_N]^2\right\}$$

$$\leqslant \frac{N}{N-1}\left\{\sum_{i=1}^{N}[c_N(i) - \bar{c}_N]^2\right\}\left\{\frac{1}{N}\sum_{i=1}^{N}[a_N(i) - \phi(u_{(i)})]^2\right\}$$

$$= \frac{N}{N-1}\left\{\sum_{i=1}^{N}[c_N(i) - \bar{c}_N]^2\right\}$$

$$\cdot \left\{E[a_N(R_1) - \phi(U_1)^2 | U^0 = u]\right\},$$

其中 $\bar{\phi}_N = \frac{1}{N}\sum_{i=1}^{N}\phi(u_i)$，$(u_{(1)} \leqslant \cdots \leqslant u_{(N)})$为$(u_1, \cdots, u_N)$按大小排列后的值. 上式两边对 U^0 取期望，得

$$E(S_N - V_N)^2 \leqslant \frac{N}{N-1}\left\{\sum_{i=1}^{N}[c_N(i) - \bar{c}_N]^2\right\}$$

$$\cdot \{E[a_N(R_1) - \phi(U_1)]^2\},$$

故由本章定理 3.6，有

$$\lim_{N\to\infty}\frac{E(S_N - V_N)^2}{\operatorname{var}V_N} \leqslant \left\{\int_0^1[\phi(u) - \bar{\phi}]^2 du\right\}^{-1}$$

$$\cdot \lim_{N\to\infty}E[a_N(R_1) - \phi(U_1)]^2$$

$$= 0.$$

因而由 Slutsky 定理，有

$$\frac{S_N - E(S_N)}{\sqrt{\operatorname{var}S_N}} = \frac{S_N - \mu_N}{\sigma_N} \quad 与 \quad \frac{V_N - E(V_N)}{\sqrt{\operatorname{var}V_N}}$$

有相同的渐近正态分布 $N(0,1)$.

例 1 考虑线性秩统计量

$$S_N = \sum_{i=1}^{N} iR_i,$$

当秩向量 $R=(R_1,\cdots,R_N)$均匀分布时，由于回归常数 $c_N(i)=i$，分值 $a_N(i)=i$ 均满足定理 4.2 之条件，而

$$E(S_N) = N\bar{c}_N\bar{a}_N = N\cdot\left(\frac{N+1}{2}\right)^2,$$

$$\begin{aligned}\operatorname{var}S_N &= \frac{1}{N-1}\left\{\sum_{i=1}^{N}[c_N(i)-\bar{c}_N]^2\right\}\left\{\sum_{i=1}^{N}[a_N(i)-\bar{a}_N]^2\right\}\\ &= \frac{1}{N-1}\left\{\sum_{i=1}^{N}\left[i-\frac{N+1}{2}\right]^2\right\}^2\\ &= (N-1)\left(\frac{N(N+1)}{12}\right)^2,\end{aligned}$$

故 $\left[S_N-N\left(\frac{N+1}{2}\right)^2\right]\Big/\left[\sqrt{N-1}\cdot\frac{N(N+1)}{12}\right]$有渐近正态分布 $N(0,1)$.

§3.5 线性符号秩统计量

定义 5.1 设 $a(1),\cdots,a(n)$是一组不全为零的非负常数，$X_1,\cdots,X_n$ 是 n 个随机变量. 随机变量 $|X_1|,\cdots,|X_n|$ 对应的秩向量为 $R^+=(R_1^+,\cdots,R_n^+)$，计数统计量 $\Psi_i=\psi(X_i)$(函数 $\psi(\cdot)$的定义参见(1.2.1)式)，则统计量

$$S^+ = \sum_{i=1}^{n}\Psi_i a(R_i^+)$$

称为**线性符号秩统计量**，常数$\{a(i)\}$称为**分值**.

例 1 取分值 $a(i)=i$ $(i=1,\cdots,n)$，对应的线性符号秩统计量即为 Wilcoxon 符号秩统计量

$$W^+ = \sum_{i=1}^{n}\Psi_i R_i^+, \tag{3.5.1}$$

取分值 $a(i)=\Phi^{-1}\left(\frac{1}{2}+\frac{1}{2}\left(\frac{i}{N+1}\right)\right)$，其中 $\Phi(x)$是标准正态分布

函数,此分值称为正态分位数分值,对应的线性符号秩统计量为

$$S^+ = \sum_{i=1}^{n} \Psi_i \Phi^{-1}\left(\frac{1}{2} + \frac{1}{2}\left(\frac{R_i^+}{N+1} \right) \right). \tag{3.5.2}$$

定理 5.1 设随机变量 $X_1, \cdots, X_n$ 相互独立同分布,分布连续且关于 0 点对称,则有线性符号秩统计量

$$S^+ = \sum_{i=1}^{n} \Psi_i a(R_i^+) \xlongequal{d} \sum_{i=1}^{n} \Psi_i a(i). \tag{3.5.3}$$

证明 统计量 S^+ 有下列表达式

$$S^+ = \sum_{i=1}^{n} \Psi_i a(R_i^+) = \sum_{i=1}^{n} \Psi_{D_i} a(i),$$

其中 $D_i = j$,当 $R_j^+ = i \ (i=1, \cdots, n)$,即 D_i 为 i 在 $R_1^+, \cdots, R_n^+$ 中的位置. 由本章第 2 节引理 2.1 知当$(R_1^+, \cdots, R_n^+)$均匀分布时,$D = (D_1, \cdots, D_n)$也在集合 $\mathscr{R}$($\mathscr{R}$ 的定义参见(3.1.6)式)上均匀分布. 而

$$\begin{aligned}
& P\{\Psi_{D_1} = \psi_1, \cdots, \Psi_{D_n} = \psi_n\} \\
&= \sum_{(d_1, \cdots, d_n)} P\{\Psi_{D_1} = \psi_1, \cdots, \Psi_{D_n} = \psi_n, D = (d_1, \cdots, d_n)\} \\
&= \sum_{(d_1, \cdots, d_n)} P\{\Psi_{d_1} = \psi_1, \cdots, \Psi_{d_n} = \psi_n, D = (d_1, \cdots, d_n)\} \\
&= \sum_{(d_1, \cdots, d_n)} P\{\Psi_{d_1} = \psi_1, \cdots, \Psi_{d_n} = \psi_n\} \cdot P\{D = (d_1, \cdots, d_n)\} \\
&= \frac{1}{n!} \sum_{(d_1, \cdots, d_n)} P\{\Psi_{d_1} = \psi_1, \cdots, \Psi_{d_n} = \psi_n\} \\
&= \frac{1}{n!} \sum_{(d_1, \cdots, d_n)} P\{\Psi_1 = \psi_1, \cdots, \Psi_n = \psi_n\} \\
&= P\{\Psi_1 = \psi_1, \cdots, \Psi_n = \psi_n\},
\end{aligned}$$

即
$$(\Psi_{D_1}, \cdots, \Psi_{D_n}) \xlongequal{d} (\Psi_1, \cdots, \Psi_n).$$

所以
$$S^+ \xlongequal{d} \sum_{i=1}^{n} \Psi_i a(i).$$

证毕.

定理 5.2 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布连续且关于 0 点对称，线性符号秩统计量 $S^+=\sum_{i=1}^{n}\Psi_i a(R_i^+)$，则

$$E(S^+)=\frac{1}{2}\sum_{i=1}^{n}a(i)=\frac{n}{2}\bar{a}, \tag{3.5.4}$$

$$\operatorname{var}(S^+)=\frac{1}{4}\sum_{i=1}^{n}a^2(i). \tag{3.5.5}$$

证明 由上一定理及 $\Psi_1,\cdots,\Psi_n$ 相互独立同分布，分布为 0,1 两点分布，概率各为 $\frac{1}{2}$，即可算得

$$E(S^+)=\sum_{i=1}^{n}a(i)E(\Psi_i)=\frac{1}{2}\sum_{i=1}^{n}a(i)=\frac{n}{2}\bar{a},$$

$$\operatorname{var}(S^+)=\sum_{i=1}^{n}a^2(i)\operatorname{var}\Psi_i=\frac{1}{4}\sum_{i=1}^{n}a^2(i).$$

证毕.

定理 5.3 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布连续且关于 0 点对称，则线性符号秩统计量 $S^+=\sum_{i=1}^{n}\Psi_i a(R_i^+)$ 的分布关于 $\frac{n}{2}\bar{a}$ 对称.

证明 对统计量 S^+，有

$$\begin{aligned}
S^+-\frac{n}{2}\bar{a} &\stackrel{d}{=}\sum_{i=1}^{n}\Psi_i a(i)-\frac{n}{2}\bar{a}\\
&\stackrel{d}{=}\sum_{i=1}^{n}(1-\Psi_i)a(i)-\frac{n}{2}\bar{a}\\
&=\frac{n}{2}\bar{a}-\sum_{i=1}^{n}\Psi_i a(i)\stackrel{d}{=}\frac{n}{2}\bar{a}-S^+.
\end{aligned}$$

证毕.

当 $X_1,\cdots,X_n$ 相互独立同分布，分布连续且关于 0 点对称时，统计量 $S^+=\sum_{i=1}^{n}\Psi_i a(R_i^+)$ 的渐近正态性（当 $n\to\infty$）的论证比较简

单.这是由于 $S^+ \overset{d}{=\!=} \sum_{i=1}^{n} \Psi_i a(i)$,而且 $\Psi_1,\cdots,\Psi_n$ 相互独立同是参数为$\frac{1}{2}$的两点分布,因而许多关于独立随机变量和的中心极限定理均可利用.下面我们利用 Liapounov 中心极限定理获得线性符号秩统计量有渐近正态分布的充分条件.

定理 5.4 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布,分布连续且关于 0 点对称,线性符号秩统计量 $S_n^+=\sum_{i=1}^{n}\Psi_i a_n(R_i^+)$,而分值 $\{a_n(i)\}$ 满足条件

$$\sum_{i=1}^{n} a_n^2(i) \Big/ \max_{1\leqslant i\leqslant n} a_n^2(i) \to \infty \quad (n\to\infty), \qquad (3.5.6)$$

则$\left(S_n^+-\frac{n}{2}\bar{a}_n\right)\Big/\left(\frac{1}{4}\sum_{i=1}^{n}a_n^2(i)\right)^{\frac{1}{2}}$有渐近正态分布 $N(0,1)$.

证明 由本节定理 5.1,只需证明统计量

$$S_n' = \sum_{i=1}^{n}\Psi_i a_n(i)$$

有渐近正态分布.S_n' 是独立随机变量的和,且

$$\begin{aligned}
&\sum_{i=1}^{n} E\left(\left|a_n(i)\Psi_i-\frac{1}{2}a_n(i)\right|^3\right) \\
&\quad = \sum_{i=1}^{n}\left[\left|\frac{1}{2}a_n(i)\right|^3\cdot\frac{1}{2}+\left|-\frac{1}{2}a_n(i)\right|^3\cdot\frac{1}{2}\right] \\
&\quad = \frac{1}{8}\sum_{i=1}^{n}|a_n(i)|^3,
\end{aligned}$$

所以

$$\begin{aligned}
\frac{\sum_{i=1}^{n} E\left(\left|a_n(i)\Psi_i-\frac{1}{2}a_n(i)\right|^3\right)}{\left(\frac{1}{4}\sum_{i=1}^{n}a_n^2(i)\right)^{3/2}} &= \frac{\sum_{i=1}^{n}|a_n(i)|^3}{\left(\sum_{i=1}^{n}a_n^2(i)\right)^{3/2}} \\
&\leqslant \frac{\left(\max_{1\leqslant i\leqslant n}a_n^2(i)\right)^{\frac{1}{2}}\cdot\sum_{i=1}^{n}a_n^2(i)}{\left(\sum_{i=1}^{n}a_n^2(i)\right)^{3/2}} = \left(\frac{\max_{1\leqslant i\leqslant n}a_n^2(i)}{\sum_{i=1}^{n}a_n^2(i)}\right)^{\frac{1}{2}} \to 0.
\end{aligned}$$

故由 Liapounov 中心极限定理知$(S_n^+-E(S_n^+))\Big/\sqrt{\mathrm{var}S_n^+}$有渐近正态分布 $N(0,1)$. 证毕.

例 2 考虑 Wilcoxon 符号秩统计量

$$W^+=\sum_{i=1}^n \Psi_i R_i^+,$$

对应的分值为 $a_n(i)=i\ (i=1,\cdots,n)$, $\max\limits_{1\leqslant i\leqslant n}a_n^2(i)=n^2$, 而 $\sum\limits_{i=1}^n a_n^2(i)=\sum\limits_{i=1}^n i^2=\dfrac{n(n+1)(2n+1)}{6}$, $\{a_n(i)\}$满足定理 5.4 的条件. 故在 H_0: 随机变量 $X_1,\cdots,X_n$ 相互独立同分布, 分布连续且关于 0 点对称时,

$$(W^+-EW^+)\Big/\sqrt{\mathrm{var}W^+}=\left(W^+-\frac{n(n+1)}{4}\right)\Big/\sqrt{\frac{n(n+1)(2n+1)}{24}}$$

有渐近正态分布 $N(0,1)$.

考虑符号统计量

$$B=\sum_{i=1}^n \Psi_i,$$

对应的分值 $a_n(i)\equiv1\ (i=1,\cdots,n)$, $\max\limits_{1\leqslant i\leqslant n}a_n^2(i)=1$, $\sum\limits_{i=1}^n a_n^2(i)=n$, $\{a_n(i)\}$满足定理 5.4 条件. 故在 H_0 下亦有$(B-E(B))\Big/\sqrt{\mathrm{var}B}=\left(B-\dfrac{n}{2}\right)\Big/\sqrt{\dfrac{n}{4}}$有渐近正态分布 $N(0,1)$.

习　题

1. 设

$$S=\sum_{i=1}^N c(i)a(R_i),\quad S'=\sum_{i=1}^N c'(i)a'(R_i)$$

是任意两个线性秩统计量, 若秩向量 $R=(R_1,\cdots,R_N)$在集合

$$\mathscr{R}=\{(1,\cdots,N)\text{ 的全部排列}\}$$

上均匀分布，证明协方差

$$\operatorname{cov}(S,S')=\frac{1}{N-1}\sum_{i=1}^{N}[c(i)-\bar{c}][c'(i)-\bar{c}']$$
$$\cdot\sum_{j=1}^{N}[a(j)-\bar{a}][a'(j)-\bar{a}'],$$

其中 $\bar{c},\bar{c}',\bar{a},\bar{a}'$ 分别是 $\{c(i)\},\{c'(i)\},\{a(i)\},\{a'(i)\}$ 的平均值.

2. 若上题中，有分值 $a(i)$ 满足

$$a(i)+a(N+1-i)=\text{常数},\quad i=1,\cdots,N,$$

而分值 $a'(i)$ 满足

$$a'(i)=a'(N+1-i),\quad i=1,\cdots,N,$$

则 S 与 S' 不相关.

3. 设 S 是一个线性秩统计量，其回归常数

$$c(i)=\begin{cases}0, & i=1,\cdots,m,\\ 1, & i=m+1,\cdots,N,\end{cases}\qquad N=m+n.$$

而分值

$$a(i)=\begin{cases}\dfrac{i}{N+1}-\dfrac{1}{4}, & i\leqslant\dfrac{N+1}{4},\\ 0, & \dfrac{N+1}{4}<i<\dfrac{3(N+1)}{4},\\ \dfrac{i}{N+1}-\dfrac{3}{4}, & i\geqslant\dfrac{3(N+1)}{4}.\end{cases}$$

证明当秩向量 $R=(R_1,\cdots,R_N)$ 在集合 $\mathscr{R}$ 上均匀分布时，S 的分布关于它的均值对称，并求出均值.

4. 考虑分值

$$a(i)=G^{-1}\left(\frac{i}{N+1}\right),$$

这里 $G(x)$ 是一个在 $[-a,a]$ 区间上取值对称分布的连续随机变量的分布函数，说明当秩向量 R 在集合 $\mathscr{R}$ 上均匀分布时，上述分值与任何一组回归常数 $\{c(i)\}$ 构成的线性秩统计量均有关于它的均值对称的分布.

5. 说明若 $N\geqslant4$，秩向量 $R=(R_1,\cdots,R_N)$ 在集合 $\mathscr{R}$ 上均匀分布，则统计量

$$S = \ln\left[\frac{R_1 \cdot R_4}{R_2 \cdot R_3}\right]$$

分布对称，并指出其对称点.

6. 说明回归常数

$$c_N(i) = \left(i - \frac{N+1}{2}\right)^2, \quad i = 1,\cdots,N$$

满足关于回归常数的 Noether 条件.

7. 找出分值

$$a_N(i) = \left(i - \frac{N+1}{2}\right)^2, \quad i = 1,\cdots,N$$

对应的分值函数，并说明它满足平方可积分值函数的条件. 从而求出线性秩统计量

$$S_N = \sum_{j=m+1}^{N}\left(R_j - \frac{N+1}{2}\right)^2$$

的极限分布，其中秩向量 $R=(R_1,\cdots,R_N)$在集合 $\mathscr{R}$ 上均匀分布.

8. 考虑线性秩统计量 $S_N=\sum_{i=1}^{N} c(i)a(R_i)$，其中秩向量 $R=(R_1,\cdots,R_N)$在集合 $\mathscr{R}$ 上均匀分布. 回归常数

$$c_N(i) = i, \quad i = 1,\cdots,N;$$

分值 $$a_N(i)=\Phi^{-1}\left(\frac{i}{N+1}\right), \quad i=1,\cdots,N,$$

其中 $\Phi(x)$是标准正态分布的分布函数. 求 S_N 的极限分布.

9. 设线性符号秩统计量

$$S^+ = \sum_{i=1}^{n}\Psi_i a(R_i^+) \quad 和 \quad S' = \sum_{i=1}^{n}\Psi_i a'(R_i^+)$$

是同一个来自连续且关于 0 点对称的分布的样本 $X_1,\cdots,X_n$ 的两个线性符号秩统计量，证明当秩向量 $R^+=(R_1^+,\cdots,R_n^+)$均匀分布时，协方差

$$\mathrm{cov}(S^+,S') = \frac{1}{4}\sum_{i=1}^{n}a(i)a'(i).$$

10. 说明定理 5.3 的条件可以减弱为：随机变量 $X_1,\cdots,X_n$ 相互独立，每一 X_i $(i=1,\cdots,n)$有一连续且关于 0 点对称的分布（不一定相同）.

第四章 功效函数

§4.1 备择假设与功效函数

非参数检验问题的功效函数由于备择假设的范围很宽，具体求得功效函数的精确表达式是很困难的，绝大部分情况基本上是不可能的. 研究功效时，常是将不同的检验作相对比较. 比较的途径基本上有两条，一是对各种典型的备择假设作统计模拟，一是求它们的渐近相对效率. 另一方面在某些情况下可在一定的限制下求最优检验，如后面将见到的局部最优秩检验.

为了建立备择假设，可以将非参数问题形式地参数化. 如考虑一样本问题：随机变量 $X_1,\cdots,X_n$ 是来自连续分布总体 $F(x-\theta)$ 的样本，函数 $F(t)$ 关于 0 点对称，检验假设

$$H_0:\ \theta=0,\qquad H_1:\ \theta>0.$$

则可确定形式参数 $\xi=(\theta,F(t))$，参数空间为

$$\Omega=\{(\theta,F(t))\,|\,\theta\geqslant 0,F(t)\text{ 连续关于 0 点对称}\},$$

H_0 所确定的空间为

$$\omega=\{(0,F(t))\,|\,F(t)\text{ 连续关于 0 点对称}\},$$

H_1 的空间为 $\Omega\backslash\omega$. 此时功效函数的定义与一般数理统计教材中的定义完全相同.

定义 1.1 对形式参数 $\xi\in\Omega$ 的有关的假设检验的**功效函数**，定义为

$$\mathscr{P}(\xi)=P_\xi\{H_0\text{ 被拒绝}\}.$$

当 $\xi\in\omega$ 时，$\mathscr{P}(\xi)$ 为第一类错误的概率，当 $\xi\in\Omega\backslash\omega$ 时，$1-\mathscr{P}(\xi)$ 为第二类错误的概率.

定义 1.2 假定形式参数 $\xi\in\Omega$,原假设和备择假设为 H_0: $\xi\in\omega$, H_1: $\xi\in\Omega\backslash\omega$,对功效函数为 $\mathscr{P}(\xi)$的一个检验,若

$$\sup_{\xi\in\omega}\mathscr{P}(\xi)\leqslant\alpha,$$

则称检验是**水平 α 的**;若对任一 $\xi\in\omega$ 均有

$$\mathscr{P}(\xi)=\alpha,$$

则称检验是**精确地水平为 α 的**.

定义 1.3 对假设检验问题 H_0: $\xi\in\omega$, H_1: $\xi\in\Omega\backslash\omega$,功效函数为 $\mathscr{P}(\xi)$的一个水平 α 的检验称为是**水平 α 无偏的**,若对一切 $\xi\in\Omega\backslash\omega$,有

$$\mathscr{P}(\xi)\geqslant\alpha.$$

定理 1.1 对一样本位置问题,设随机变量 $X_1,\cdots,X_n$ 相互独立同分布,分布为 $F(x-\theta)$,函数 $F(t)$连续且关于 0 点对称,检验假设

$$H_0:\theta=\theta_0,\qquad H_1:\theta>\theta_0.$$

检验统计量为 $S(X_1,\cdots,X_n)$,否定域的形式为

$$S(X_1,\cdots,X_n)\geqslant c,\tag{4.1.1}$$

这里 c 是否定域的临界值,而 S 具有性质:对一切 $x_1,\cdots,x_n$ 及 $k\geqslant0$,有

$$S(x_1+k,\cdots,x_n+k)\geqslant S(x_1,\cdots,x_n),\tag{4.1.2}$$

则这个检验有对 θ 单调的功效函数,即当 $\theta\leqslant\theta'$ 及一切连续对称分布 $F(x)$,功效函数 $\mathscr{P}_S(\cdot)$满足

$$\mathscr{P}_S(\theta,F)\leqslant\mathscr{P}_S(\theta',F).\tag{4.1.3}$$

证明 随机变量 $X_1,\cdots,X_n$ 相互独立同分布,分布函数为 $F(x-\theta)$,平移后,随机变量 $X_1+(\theta'-\theta),\cdots,X_n+(\theta'-\theta)$相互独立同分布,分布函数为 $F(x-\theta')$. 因此,当 $\theta'-\theta\geqslant0$ 时,功效函数

$$\begin{aligned}\mathscr{P}_S(\theta,F)&=P\{S(X_1,\cdots,X_n)\geqslant c\,|\,X_i\sim F(x-\theta),i=1,\cdots,n\}\\&\leqslant P\{S(X_1+(\theta'-\theta),\cdots,X_n+(\theta'-\theta))\geqslant c\\&\qquad|\,X_i\sim F(x-\theta),i=1,\cdots,n\}\end{aligned}$$

$$= P\{S(X_1 + (\theta' - \theta), \cdots, X_n + (\theta' - \theta)) \geqslant c$$

$$|X_i + (\theta' - \theta) \sim F(x - \theta'), i = 1, \cdots, n\}$$

$$= P\{S(X_1^*, \cdots, X_n^*) \geqslant c | X_i^* \sim F(x - \theta'), i = 1, \cdots, n\}$$

$$= \mathscr{P}_S(\theta', F).$$

证毕.

定理 1.1′ 定理 1.1 中,若否定域为

$$S(X_1, \cdots, X_n) \leqslant c, \tag{4.1.4}$$

而检验统计量 S 有性质:对一切 $x_1, \cdots, x_n$ 及 $k \geqslant 0$,有

$$S(x_1 + k, \cdots, x_n + k) \leqslant S(x_1, \cdots, x_n), \tag{4.1.5}$$

则仍有定理 1.1 的结论.

例 1 一样本问题.设随机变量 $X_1, \cdots, X_n$ 相互独立同分布,分布为 $F(x-\theta)$,函数 $F(t)$ 连续且关于 0 点对称.检验假设

$$H_0: \theta = \theta_0, \qquad H_1: \theta > \theta_0.$$

可用下列一些检验统计量进行检验.

(1) ***T* 检验** 统计量

$$T = \frac{\overline{X} - \theta_0}{S/\sqrt{n}}, \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

否定域为 $T \geqslant t_{n-1}(\alpha)$ ($t_{n-1}(\alpha)$是显著性水平为 α 的临界值).

(2) **符号检验**(参见第一章第 2 节) 统计量

$$B = \sum_{i=1}^{n} \psi(X_i - \theta_0),$$

否定域为 $B \geqslant b(n, \alpha)$ ($b(n, \alpha)$是显著性水平为 α 的临界值).

(3) **符号秩检验**(参见第一章第 4 节) 统计量

$$W^+ = \sum_{i=1}^{n} \psi(X_i - \theta_0) R_i^+ = \sum_{i=1}^{n} \sum_{j \leqslant i} \psi(X_i + X_j - 2\theta_0),$$

否定域为 $W^+ \geqslant \omega(n, \alpha)$ ($\omega(n, \alpha)$是显著性水平为 α 的临界值).

上述这三个统计量均满足定理 1.1 的条件.对任意 $x_1, \cdots, x_n$ 及 $k > 0$,有

$$T(x_1 + k, \cdots, x_n + k) = \frac{\overline{x} + k - \theta_0}{S/\sqrt{n}}$$

$$= T(x_1,\cdots,x_n) + \frac{k}{S/\sqrt{n}} \geqslant T(x_1,\cdots,x_n),$$

$$B(x_1+k,\cdots,x_n+k) = \sum_{i=1}^{n}\psi(x_i+k-\theta_0)$$

$$\geqslant \sum_{i=1}^{n}\psi(x_i-\theta_0) = B(x_1,\cdots,x_n),$$

$$W^+(x_1+k,\cdots,x_n+k) = \sum_{i=1}^{n}\sum_{j<i}\psi(x_i+x_j+2k-2\theta_0)$$

$$\geqslant \sum_{i=1}^{n}\sum_{j\leqslant i}\psi(x_i+x_j-2\theta_0) = W^+(x_1,\cdots,x_n).$$

所以这三个检验都有对 θ 单调的功效函数.

虽然 T,B,W^+都有对 θ 单调的功效函数,符号检验、符号秩检验对一切连续关于 θ_0 对称的分布组成的分布类是适应任意分布的.对这一分布类中的任一分布 $F(x)$,检验水平是相同的.因而当 $\theta>\theta_0$ 时,对 H_0 下的一点 $\xi_0=(\theta_0,F_0)$以及 H_1 下的一点 $\xi=(\theta,F_1)$,有

$$\mathscr{P}(\theta_0,F_0) = \mathscr{P}(\theta_0,F_1) \leqslant \mathscr{P}(\theta,F_1).$$

因而它们是水平无偏的.然而 T 检验对一切连续关于 θ_0 对称的分布组成的分布类,不是适应任意分布的.其检验水平依赖于总体分布 $F(x)$.事实上,当 $F(x)$为 Cauchy 分布时的检验水平远小于 $F(x)$为正态分布时的检验水平.因而 T 虽然有对 θ 单调的功效函数,但可找到 $\theta>\theta_0$ 及 F_0 为正态分布,F_1 为 Cauchy 分布,使 H_0 下的一点 $\xi_0=(\theta_0,F_0)$及 H_1 下的一点 $\xi=(\theta,F_1)$,仍有

$$\mathscr{P}(\theta_0,F_0) > \mathscr{P}(\theta,F_1),$$

所以 T 检验在连续关于 θ_0 对称的分布类中不是水平无偏的.

对二样本问题有类似的定理.

定理 1.2 设随机变量 $X_1,\cdots,X_m$ 相互独立同分布,分布函数 $F(x)$连续;随机变量 $Y_1,\cdots,Y_n$ 相互独立同分布,分布函数为 $F(x-\Delta)$.检验假设

$$H_0: \Delta = 0, \qquad H_1: \Delta > 0.$$

对上述二样本检验问题有检验统计量 $S(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$，否定域为

$$S(X_1,\cdots,X_m;Y_1,\cdots,Y_n) \geqslant c, \qquad (4.1.6)$$

这里 c 是显著性水平为 α 的临界值，而 S 具有性质：对任一 $x_1,\cdots,x_m;y_1,\cdots,y_n$ 及 $k>0$，有

$$S(x_1,\cdots,x_m;y_1+k,\cdots,y_n+k) \geqslant S(x_1,\cdots,x_m;y_1,\cdots,y_n), \qquad (4.1.7)$$

则这个检验有对 Δ 单调的功效函数，即对一切连续分布 $F(x)$ 及 $\Delta \leqslant \Delta'$，有

$$\mathscr{P}_S(\Delta,F) \leqslant \mathscr{P}_S(\Delta',F). \qquad (4.1.8)$$

证明 与定理 1.1 类似.

例 2 Wilcoxon 秩和统计量(参见第一章第 3 节)

$$W = \sum_{i=1}^{n} R_i = \sum_{i=1}^{m}\sum_{j=1}^{n}\psi(Y_j - X_i) + \frac{n(n+1)}{2}$$

有

$$\begin{aligned} & W(x_1,\cdots,x_m;y_1+k,\cdots,y_n+k) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\psi(y_j+k-x_i) + \frac{n(n+1)}{2} \\ &\geqslant \sum_{i=1}^{m}\sum_{j=1}^{n}\psi(y_j-x_i) + \frac{n(n+1)}{2} \\ &= W(x_1,\cdots,x_m;y_1,\cdots,y_n), \end{aligned}$$

所以 W 对 Δ 有单调的功效函数. W 统计量对连续分布类是适应任意分布的，因而 W 也是水平无偏的.

当考虑检验的大样本性质时，首先碰到相合性的概念.

定义 1.4 以检验统计量序列 $\{T_n\}$ 记假设检验问题 $H_0: \xi \in \omega$, $H_1: \xi \in \Omega\backslash\omega$ 的一个水平 α 的检验序列，检验序列 $\{T_n\}$ 称为对备择集合 $\Omega\backslash\omega$ 是**相合的**，如果对任意一个 $\xi \in \Omega\backslash\omega$，有

$$\lim_{n\to\infty}\mathscr{P}_{T_n}(\xi) = 1.$$

定理 1.3 以检验统计量序列 $\{T_n\}$ 记假设检验问题 $H_0: \xi \in$

ω, H_1: $\xi\in\Omega\backslash\omega$ 的一列水平 α 的检验统计量，否定域有形式 $T_n\geqslant c_n$，这里 c_n 是显著性水平为 α 的临界值. 若存在函数 $K(\xi)$，使得 ξ 为真时，

$$T_n \xrightarrow{P} K(\xi), \quad 对任一\ \xi\in\Omega,$$

且当 $\xi\in\omega$ 时，$K(\xi)=k_0$（k_0 为一常数）；当 $\xi\in\Omega\backslash\omega$ 时，$K(\xi)>k_0$. 又对临界值 c_n，有

$$\lim_{n\to\infty} c_n \leqslant k_0,$$

则 $\{T_n\}$ 检验序列对备择集合 $\Omega\backslash\omega$ 是相合的.

证明 对任一 $\xi\in\Omega\backslash\omega$，令

$$\varepsilon=\frac{K(\xi)-k_0}{2}>0.$$

当 n 充分大时，可以有 $c_n\leqslant k_0+\varepsilon=K(\xi)-\varepsilon$，因此

$$\lim_{n\to\infty}\mathscr{P}_{T_n}(\xi)=\lim_{n\to\infty}P_\xi\{T_n\geqslant c_n\}\geqslant\lim_{n\to\infty}P_\xi\{T_n\geqslant K(\xi)-\varepsilon\}$$

$$\geqslant\lim_{n\to\infty}P_\xi\{|T_n-K(\xi)|\leqslant\varepsilon\}=1.$$

所以

$$\lim_{n\to\infty}\mathscr{P}_{T_n}(\xi)=1,$$

即 $\{T_n\}$ 对 $\Omega\backslash\omega$ 是相合的. 证毕.

例 3 一样本问题. 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布为 $F(x-\theta)$，函数 $F(t)$ 连续且关于 0 点对称. 检验假设

$$H_0:\ \theta=0, \qquad H_1:\ \theta>0.$$

考虑符号秩检验，统计量 $\dfrac{1}{\binom{n}{2}}W^+=\dfrac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{j\leqslant i}\psi(X_i+X_j)$. 当 $n\to\infty$ 时，对一切 $\theta\geqslant0$ 及关于 θ 对称的连续分布 $F(x)$，$\xi=(\theta,F)$ 均有

$$\frac{1}{\binom{n}{2}}W^+\xrightarrow{P}P_\xi\{X_1+X_2>0\}\overset{\text{def}}{=\!=}K(\xi).$$

又对 $\xi_0=(0,F)$，

$$P_{\xi_0}\{X_1 + X_2 > 0\} = \iint\limits_{x_1+x_2>0} \mathrm{d}F(x_1)\mathrm{d}F(x_2)$$
$$= \int_{-\infty}^{+\infty}\left[\int_{-x_2}^{+\infty}\mathrm{d}F(x_1)\right]\mathrm{d}F(x_2)$$
$$= \int_{-\infty}^{+\infty}[1 - F(-x_2)]\mathrm{d}F(x_2)$$
$$= \int_{-\infty}^{+\infty}F(x_2)\mathrm{d}F(x_2) = \frac{1}{2},$$

对 $\xi=(\theta,F)$，$\theta>0$，$F(x)$关于 θ 对称，

$$P_{\xi}\{X_1 + X_2 > 0\} > P_{\xi}\{X_1 + X_2 - 2\theta > 0\}$$
$$= P_{\xi}\{(X_1 - \theta) + (X_2 - \theta) > 0\} = \frac{1}{2}.$$

又在 H_0：$\theta=0$ 下，

$$\frac{W^+ - E(W^+)}{\sqrt{\mathrm{var}W^+}} \text{ 有渐近分布 } N(0,1),$$

故否定域(以 $\omega(n,\alpha)$记显著性水平为 α 的临界值)

$$\frac{1}{\binom{n}{2}}W^+ \geqslant \omega(n,\alpha)$$

等价于

$$\frac{W^+ - E(W^+)}{\sqrt{\mathrm{var}W^+}} \geqslant \frac{\omega(n,\alpha) - \frac{1}{\binom{n}{2}}E(W^+)}{\frac{1}{\binom{n}{2}}\sqrt{\mathrm{var}W^+}},$$

应有

$$\frac{\binom{n}{2}\omega(n,\alpha) - E(W^+)}{\sqrt{\mathrm{var}W^+}} \to z_\alpha,$$

其中 z_α 为标准正态分布的 $1-\alpha$ 分位数，即有

$$\frac{\frac{n(n-1)}{2}\omega(n,\alpha)-\frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\to z_\alpha,$$

故对临界值 $\omega(n,\alpha)$，有

$$\lim_{n\to\infty}\omega(n,\alpha)=\lim_{n\to\infty}\frac{\frac{n(n+1)}{4}+z_\alpha\sqrt{\frac{n(n+1)(2n+1)}{24}}}{\frac{n(n-1)}{2}}=\frac{1}{2}.$$

定理 1.3 的条件均满足，故检验对 $\theta>0$ 是相合的.

例 4 二样本问题. 设随机变量 $X_1,\cdots,X_m$ 相互独立同分布，分布函数 $F(x)$ 连续；随机变量 $Y_1,\cdots,Y_n$ 相互独立同分布，分布函数为 $F(x-\Delta)$. 检验假设

$$H_0:\ \Delta=0,\qquad H_1:\ \Delta>0.$$

记 $\xi=(\Delta,F)$，$\Omega=\{(\Delta,F)\,|\,\Delta\geqslant 0,F(x)\text{连续}\}$，$\omega=\{(0,F)\,|\,F(x)\text{连续}\}$. 检验统计量

$$U=\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(Y_j-X_i).$$

对任意 $\xi\in\Omega$，有

$$U\xrightarrow{P}E(U)=P_\xi\{X<Y\}\stackrel{\text{def}}{=\!=}K(\xi)$$

$\left(\text{当 } N=m+n\to\infty,\frac{m}{N}\to\lambda,0<\lambda<1\right)$，当 $\xi\in\omega$ 时，$K(\xi)=\frac{1}{2}$，$\xi\in\Omega\backslash\omega$ 时，$K(\xi)>\frac{1}{2}$.

在 H_0 下，$[U-E(U)]/\sqrt{\mathrm{var}U}$ 有渐近分布 $N(0,1)$. 而 $E(U)=\frac{1}{2}$，$\mathrm{var}U=\frac{m+n+1}{12nm}$. 水平 α 的否定域

$$U\geqslant c_N\quad(c_N\text{ 为相应的临界值}),$$

则应有

$$[c_N-E(U)]/\sqrt{\mathrm{var}U}=\left[c_N-\frac{1}{2}\right]\Big/\sqrt{\frac{m+n+1}{12mn}}\to z_\alpha,$$

其中 z_α 为标准正态分布的 $1-\alpha$ 分位数. 故有

$$\lim_{N\to\infty} c_N = \lim_{N\to\infty}\left[\frac{1}{2} + z_\alpha\sqrt{\frac{m+n+1}{12mn}}\right] = \frac{1}{2}.$$

满足定理 1.3 的各项条件，因此 U 检验对 $\Delta>0$ 是相合的. Wilcoxon 秩和检验 W 与上述检验是等价的，所以也是相合的.

定理 1.3′ 以检验统计量序列 $\{T_n\}$ 记假设检验问题. H_0: $\xi\in\omega$, H_1: $\xi\in\Omega\backslash\omega$ 的一列水平 α 的检验统计量，否定域为 $T_n\leqslant c_n$ (c_n 是显著性水平为 α 的临界值)，若存在函数 $K(\xi)$，使得 ξ 为真时，

$$T_n \xrightarrow{P} K(\xi), \quad 对任一 \xi\in\Omega,$$

且当 $\xi\in\omega$ 时，$K(\xi)=k_0$；当 $\xi\in\Omega\backslash\omega$ 时，$K(\xi)<k_0$. 又对临界值 c_n，有

$$\lim_{n\to\infty} c_n \geqslant k_0,$$

则 $\{T_n\}$ 对备择集合 $\Omega\backslash\omega$ 是相合的.

证明 类似定理 1.3.

§ 4.2 备择假设的其他提法

在非参数检验问题中，除上一节所述采用形式化参数的提法外，也可见到一些其他的提法. 下面介绍两种提法.

1. 纯自然的提法

例如二样本问题：设随机变量 $X_1,\cdots,X_m$ 相互独立同分布，分布函数 $F(x)$ 连续；随机变量 $Y_1,\cdots,Y_n$ 相互独立同分布，分布函数 $G(x)$ 连续. 检验假设

H_0: $F(x)=G(x)$;

H_1: $F(x)$ 随机地小于 $G(x)$，即 $F(x)\geqslant G(x)$，且对某些 x 值不等号成立.

对上述提法的检验问题也有类似上节定理 1.2 的定理.

定理 2.1 对上述检验(H_0,H_1)问题,检验统计量为$S(X_1,\cdots X_m;Y_1,\cdots,Y_n)$,否定域为

$$S(X_1,\cdots,X_m;Y_1,\cdots,Y_n)\geqslant c\ (c\text{ 为显著性水平 }\alpha\text{ 的临界值}).$$

而统计量 S 满足:对任一组 $x_1,\cdots,x_m$ 和 $y_1\leqslant y_1',\cdots,y_n\leqslant y_n'$,有

$$S(x_1,\cdots,x_m;y_1',\cdots,y_n')\geqslant S(x_1,\cdots,x_m;y_1,\cdots,y_n),$$

那么基于统计量 S 的检验是单调的,即如果 $G(x),H(x)$ 是两个连续分布,而 $G(x)$ 随机地小于 $H(x)$,则有

$$\mathscr{P}_S(F,G)\leqslant\mathscr{P}_S(F,H).$$

证明 由 $G(x)\geqslant H(x)$ 推知下列集合包含关系

$$\{x|G(x)\geqslant t\}\supset\{x|H(x)\geqslant t\}.$$

从而对分布函数 $H(t),G(t)$ 的 t 分位数,有

$$H^{-1}(t)=\inf\{x|H(x)\geqslant t\}\geqslant G^{-1}(t)=\inf\{x|G(x)\geqslant t\}.$$

又对连续分布 $G(x),H(x)$ 有:若随机变量 $Y\sim G(x)$,则随机变量 $G(Y)$ 遵从$(0,1)$均匀分布,反之若随机变量 U 遵从$(0,1)$均匀分布,则随机变量 $G^{-1}(U)\sim G(x)$. 对 $H(x)$ 有类似结论:随机变量 $H^{-1}(U)\sim H(x)$. 所以

$$\begin{aligned}\mathscr{P}_S(F,G)&=P\{S(X_1,\cdots,X_m;Y_1,\cdots,Y_n)\geqslant c|\ X_i\sim F(x)\\&\qquad(i=1,\cdots,m),Y_j\sim G(x)\ (j=1,\cdots,n)\}\\&=P\{S(X_1,\cdots,X_m;G^{-1}(U_1),\cdots,G^{-1}(U_n))\geqslant c|\ X_i\sim F(x)\\&\qquad(i=1,\cdots,m),U_j\sim(0,1)\text{ 均匀}(j=1,\cdots,n)\}\\&\leqslant P\{S(X_1,\cdots,X_m;H^{-1}(U_1),\cdots,H^{-1}(U_n))\geqslant c|\ X_i\sim F(x)\\&\qquad(i=1,\cdots,m),U_j\sim(0,1)\text{ 均匀}(j=1,\cdots,n)\}\\&=P\{S(X_1,\cdots,X_m;Y_1,\cdots,Y_n)\geqslant c|\ X_i\sim F(x)\\&\qquad(i=1,\cdots,m),Y_j\sim H(x)(j=1,\cdots,n)\}\\&=\mathscr{P}_S(F,H).\end{aligned}$$

证毕.

2. 统计学家 Lehmann 的提法

统计学家 Lehmann 曾提出一类非参数假设，有针对性地提出范围窄一点的备择假设. 其一般性的提法如下：设 $G_1(u)$ 和 $G_2(u)$ 是定义在 $[0,1]$ 上的单调非降连续函数，且 $u=0$ 时，其值为 0，$u=1$ 时，其值为 1. 则对任一连续分布函数 $F(x)$，$G_1(F(x))$ 和 $G_2(F(x))$ 仍然是分布函数. 假定：随机变量 $X_1,\cdots,X_m$ 为来自 $G_1(F(x))$ 的简单随机样本；随机变量 $Y_1,\cdots,Y_n$ 为来自 $G_2(F(x))$ 的简单随机样本. 检验假设

H_0: $G_1(u)=G_2(u)$，对一切 $u\in(0,1)$；

H_1: $G_1(u)$ 随机地小于 $G_2(u)$.

通常 G_1 和 G_2 间有一个确定的参数联系着. 例如

$$G_1(u) = u, \quad G_2(u) = u^K, \quad K \text{ 正整数},$$

则假设检验问题变成：

$$H_0: K = 1, \qquad H_1: K > 1.$$

这一假设 H_0 是说 X_i 和 Y_j 是来自同一总体的样本，而 H_1 是说 Y_j 不是简单的与 X_i 同样的样本，而是 K 个独立随机变量的最大值，即 $X_i\sim F(x)$，$Y_j\sim F^K(x)$. 又如

$$G_1(u) = 1 - (1-u)^{1+K}, \quad G_2(u) = u^{1+K}, \quad K \text{ 非负整数},$$

则假设检验问题变成

$$H_0: K = 0, \qquad H_1: K > 0.$$

这一假设 H_0 是说 X_i 和 Y_j 是来自同一总体 $F(x)$ 的样本，而 H_1 是说 X_i 是 $1+K$ 个独立变量的最小值，Y_j 是 $1+K$ 个独立变量的最大值.

Lehmann 这种提法有一定实际背景. 在理论上其重要点是 $X_1,\cdots,X_m;Y_1,\cdots,Y_n$ 的混合样本对应的秩 $R=(Q_1,\cdots,Q_m;R_1,\cdots,R_n)$ 的分布仅依赖 $G_1(x)$，$G_2(x)$，而不依赖于 $F(x)$. 因而秩统计量在 H_1 下也是非参数适应任意分布的. 它们的功效函数与 $F(x)$ 无关.

定理 2.2 设随机变量 $X_1,\cdots,X_m$ 相互独立同分布，分布为 $G(F(x))$，有密度 $g_1(F(x))f(x)$；随机变量 $Y_1,\cdots,Y_n$ 相互独立同分布，分布为 $G_2(F(x))$，有密度 $g_2(F(x))f(x)$，其中 $f(x)$是分布 $F(x)$的密度，$g_1(t),g_2(t)$分别是 $G_1(t),G_2(t)$的密度. 以 $R=(Q_1,\cdots,Q_m,R_1,\cdots,R_n)$记混合样本 $X_1,\cdots,X_m,Y_1,\cdots,Y_n$ 对应的秩向量，则对$(1,\cdots,m+n)$的任意一个排列 $r=(q_1,\cdots,q_m,r_1,\cdots,r_n)$，事件$\{R=r\}$的概率仅依赖于 $G_1(x),G_2(x)$，而与$F(x)$无关.

证明 令集合

$$A=\{(x_1,\cdots,x_m,y_1,\cdots,y_n)|x_i\text{ 有秩 }q_i(i=1,\cdots,m),\ y_j\text{ 有秩 }r_j(j=1,\cdots,n)\},\tag{4.2.1}$$

则

$$P\{R=r\}=\int\cdots\int_A\Big[\prod_{i=1}^m g_1(F(x_i))\Big]\Big[\prod_{j=1}^n g_2(F(y_j))\Big]\times\prod_{i=1}^m f(x_i)\prod_{j=1}^n f(y_j)\mathrm{d}x_1\cdots\mathrm{d}x_m\mathrm{d}y_1\cdots\mathrm{d}y_n.$$

作变换 $t_k=F(x_k),k=1,\cdots,m,t_k=F(y_{k-m}),k=m+1,\cdots,m+n$，则上述概率

$$P\{R=r\}=\int\cdots\int_{A^*}\Big[\prod_{i=1}^m g_1(t_i)\Big]\Big[\prod_{j=1}^n g_2(t_{m+j})\Big]\mathrm{d}t_1\cdots\mathrm{d}t_{m+n},$$

其中 A^*是 A 在上述变换下对应的 $t_1,\cdots,t_{m+n}$的集合，由于 $F(x)$连续单调，故

$$A^*=\{(t_1,\cdots,t_{m+n})|0\leqslant t_i\leqslant 1\ (i=1,\cdots,m+n)\text{ 且 }(t_1,\cdots,t_{m+n})\text{ 对应的秩为}(q_1,\cdots,q_m,r_1,\cdots,r_n)\},\tag{4.2.2}$$

此集合与 $F(X)$无关. 故从上面概率 $P\{R=r\}$的表达式看出定理结论成立. 证毕.

本节的最后我们给出一样本问题和二样本问题在备择假设下，秩统计量的功效函数的两个定理.

定理 2.3 设随机变量 $X_1,\cdots,X_m$ 相互独立同分布，分布为

$F_1(x)$，有密度 $f_1(x)$；随机变量 $Y_1,\cdots,Y_n$ 相互独立同分布，分布为 $F_2(x)$，有密度 $f_2(x)$. 又分布 $H(x)$的密度为 $h(x)$，$h(x)$的不为零的区域包含 $f_1(x)$和 $f_2(x)$的不为零的区域，则对混合样本的秩向量 $R=(Q_1,\cdots,Q_m,R_1,\cdots,R_n)$，有概率分布

$$P\{R=r\}=\frac{1}{N!}E\left[\frac{\prod_{i=1}^{m}f_1(V_{(q_i)})\prod_{j=1}^{n}f_2(V_{(r_j)})}{\prod_{i=1}^{N}h(V_{(i)})}\right], \quad (4.2.3)$$

其中 $r=(q_1,\cdots,q_m,r_1,\cdots,r_n)$是$(1,\cdots,m+n)$的一个排列. $N=m+n$. $V_{(1)}<\cdots<V_{(N)}$是 $H(x)$的 N 个独立样本的次序统计量.

证明 令集合

$$A=\{(x_1,\cdots,x_m,y_1,\cdots,y_n)\mid x_i \text{ 有秩 } q_i(i=1,\cdots,m),$$
$$y_j \text{ 有秩 } r_j(j=1,\cdots,n)\},$$

则概率分布

$$P\{R=r\}$$
$$=\int\cdots\int_A\left[\prod_{i=1}^{m}f_1(x_i)\right]\left[\prod_{j=1}^{n}f_2(y_j)\right]\mathrm{d}x_1\cdots\mathrm{d}x_m\mathrm{d}y_1\cdots\mathrm{d}y_n.$$

若记 $x_i=t_{(q_i)}(i=1,\cdots,m)$，$y_j=t_{(r_j)}(j=1,\cdots,n)$，则

$$P\{R=r\}=\int\cdots\int_{t_{(1)}<\cdots<t_{(N)}}\left[\prod_{i=1}^{m}f_1(t_{(q_i)})\right]\left[\prod_{j=1}^{n}f_2(t_{(r_j)})\right]\mathrm{d}t_{(1)}\cdots\mathrm{d}t_{(N)}$$

$$=\frac{1}{N!}\int\cdots\int_{t_{(1)}<\cdots<t_{(N)}}\left[\frac{\prod_{i=1}^{m}f_1(t_{(q_i)})\prod_{j=1}^{n}f_2(t_{(r_j)})}{\prod_{i=1}^{N}h(t_{(i)})}\right]$$
$$\cdot N!\prod_{i=1}^{N}h(t_{(i)})\mathrm{d}t_{(1)}\cdots\mathrm{d}t_{(N)}$$

$$=\frac{1}{N!}E\left[\frac{\prod_{i=1}^{m}f_1(V_{(q_i)})\prod_{j=1}^{n}f_2(V_{(r_j)})}{\prod_{i=1}^{N}h(V_{(i)})}\right].$$

证毕.

例 1 利用定理可计算前述 $G_1(u)=u$, $G_2(u)=u^K$（K 正整数）情形下秩统计量的功效函数. 由定理 2.2 秩向量的分布不依赖 $F(x)$，我们可以假定 $F(x)$ 为 $(0,1)$ 上均匀分布. 即随机变量 $X_1, \cdots, X_m$ 相互独立同分布，分布 $G_1(x)=x$，密度 $g_1(x)=1$ $(0\leqslant x\leqslant 1)$；随机变量 $Y_1,\cdots,Y_n$ 相互独立同分布，分布 $G_2(x)=x^K$，密度 $g_2(x)=Kx^{K-1}$ $(0\leqslant x\leqslant 1)$. 又取 $H(x)$ 为 $[0,1]$ 上均匀分布，则由定理 2.3 得

$$P\{R=r\}=\frac{K^n}{N!}E\Big[\prod_{j=1}^{n}\{V_{(r_j)}\}^{K-1}\Big], \tag{4.2.4}$$

其中 $V_{(1)}<\cdots<V_{(N)}$ 是分布 $H(x)$，即 $[0,1]$ 均匀分布样本的次序统计量. 今将 $r_1,\cdots,r_n$ 按值的大小排列后记为 $s_1,\cdots,s_n$，则 $V_{(s_1)},\cdots,V_{(s_n)}$ 的联合密度为（令 $s_0=0, s_{n+1}=N+1$）

$$u(t_1,\cdots,t_n)=\begin{cases}\dfrac{N!}{\prod\limits_{i=0}^{n}(s_{i+1}-s_i-1)!}\prod\limits_{j=0}^{n}(t_{j+1}-t_j)^{s_{j+1}-s_j-1}, \\ \qquad 0\leqslant t_1<\cdots<t_n\leqslant 1, \\ 0, \quad 其他.\end{cases}$$

若令 $Z_i=V_{(s_i)}/V_{(s_{i+1})}$ $(i=1,\cdots,n-1)$，$Z_n=V_{(s_n)}$，则随机变量 $Z_1,\cdots,Z_n$ 的联合密度为

$$l(z_1,\cdots,z_n)=\begin{cases}\dfrac{N!}{\prod\limits_{i=0}^{n}(s_{i+1}-s_i-1)!}\prod\limits_{j=1}^{n}z_j^{s_j-1}(1-z_j)^{s_{j+1}-s_j-1}, \\ \qquad 0\leqslant z_j\leqslant 1, j=1,\cdots,n, \\ 0, \quad 其他.\end{cases}$$

因而 $Z_1,\cdots,Z_n$ 是相互独立的且 Z_j 有贝塔分布 $\mathrm{B}(s_j, s_{j+1}-s_j)$. 由于

$$\prod_{j=1}^{n} V_{(r_j)} = Z_1 Z_2^2 \cdots Z_n^n,$$

所以概率

$$\begin{aligned} P\{R = r\} &= \frac{K^n}{N!} \prod_{j=1}^{n} E[Z_j^{j(K-1)}] \\ &= \frac{K^n}{\Gamma(s_1)} \prod_{j=1}^{n} \left[\frac{\Gamma(s_j + jK - j)}{\Gamma(s_{j+1} + jK - j)} \right]. \end{aligned} \tag{4.2.5}$$

定理 2.4 设随机变量 $X_1, \cdots, X_n$ 相互独立同分布，分布函数 $F(x)$ 连续，$R^+ = (R_1^+, \cdots, R_n^+)$ 是 $X_1, \cdots, X_n$ 对应的绝对秩向量，$\Psi_i = \psi(X_i)$ 为 X_i 对应的计数统计量，则对任一个 $d = (d_1, \cdots, d_n)$ 是分量为 0 或 1 的向量及 $r^+ = (r_1^+, \cdots, r_n^+)$ 是 $(1, \cdots, n)$ 的一个排列，有联合概率分布

$$P\{\Psi = d, R^+ = r^+\}$$

$$= [F(0)]^n \left[\frac{1 - F(0)}{F(0)} \right]^{\sum_{i=1}^{n} d_i} \cdot \frac{1}{n!} E \left[\frac{\prod_{i=1}^{n} f(V_{(r_i^+)} | d_i)}{\prod_{i=1}^{n} h(V_{(i)})} \right], \tag{4.2.6}$$

其中 $\Psi = (\Psi_1, \cdots, \Psi_n)$，$f(x|1)$ 是 X_i 取正值的条件下的条件密度：

$$f(x|1) = \begin{cases} 0, & t \leqslant 0, \\ \dfrac{f(t)}{1 - F(0)}, & t > 0; \end{cases} \tag{4.2.7}$$

$f(x|0)$ 是 X_i 取负值的条件下 $|X_i|$ 的条件密度：

$$f(x|0) = \begin{cases} 0, & t \leqslant 0, \\ \dfrac{f(-t)}{F(0)}, & t > 0, \end{cases} \tag{4.2.8}$$

$V_{(1)} < \cdots < V_{(n)}$ 是密度为 $h(x)$ 的独立样本的次序统计量，$h(x)$ 不为 0 的区域包含 $f(x|1)$ 和 $f(x|0)$ 不为 0 的区域.

证明 令

$$\delta_i = \begin{cases} 1, & X_i > 0, \\ -1, & X_i \leqslant 0, \end{cases} \quad i = 1, \cdots, n,$$

则联合概率分布

$$\begin{aligned} & P\{\Psi = d, R^+ = r^+\} \\ & \quad = P\{\Psi = d, (\delta_1 X_1, \cdots, \delta_n X_n) \text{ 有秩 } r^+\} \\ & \quad = P\{\Psi = d\} \cdot P\{(\delta_1 X_1, \cdots, \delta_n X_n) \text{ 有秩 } r^+ \mid \Psi = d\} \end{aligned}$$

有
$$P\{\Psi = d\} = [F(0)]^n \left[\frac{1 - F(0)}{F(0)}\right]^{\sum_{i=1}^n d_i}.$$

而

$$\begin{aligned} & P\{\delta_i X_i \leqslant x \mid \Psi_i = 0\} \\ & = \begin{cases} 0, & \text{当 } x \leqslant 0 \text{ 时}, \\ \dfrac{P\{\delta_i X_i \leqslant x, \Psi_i = 0\}}{P\{\Psi_i = 0\}} = \dfrac{P\{-x \leqslant X_i \leqslant 0\}}{P\{X_i \leqslant 0\}} & \\ \qquad = \dfrac{F(0) - F(-x)}{F(0)}, & \text{当 } x > 0 \text{ 时}. \end{cases} \end{aligned}$$

对应的密度为

$$f(x|0) = \begin{cases} 0, & x \leqslant 0, \\ \dfrac{f(-x)}{F(0)}, & x > 0; \end{cases}$$

$$\begin{aligned} & P\{\delta_i X_i \leqslant x \mid \Psi_i = 1\} \\ & = \begin{cases} 0, & x \leqslant 0, \\ \dfrac{P\{0 < X_i \leqslant x\}}{P\{X_i > 0\}} = \dfrac{F(x) - F(0)}{1 - F(0)}, & x > 0. \end{cases} \end{aligned}$$

对应的密度为

$$f(x|1) = \begin{cases} 0, & x \leqslant 0, \\ \dfrac{f(x)}{1 - F(0)}, & x > 0, \end{cases}$$

所以

$$P\{(\delta_1 X_1, \cdots, \delta_n X_n) \text{ 有秩 } r^+ \mid \Psi = d\}$$

$$= \int \cdots \int_{\left\{\substack{(x_1,\cdots,x_n)\\ 有秩 r^+}\right\}} \prod_{i=1}^{n} f(x_i|d_i)\mathrm{d}x_1\cdots\mathrm{d}x_n$$

$$= \int \cdots \int_{t_{(1)}<\cdots<t_{(n)}} \prod_{i=1}^{n} f(t_{(r_i^+)}|d_i)\mathrm{d}t_{(1)}\cdots\mathrm{d}t_{(n)}$$

$$= \frac{1}{n!}\int \cdots \int_{t_{(1)}<\cdots<t_{(n)}} \frac{\prod_{i=1}^{n} f(t_{(r_i^+)}|d_i)}{\prod_{i=1}^{n} h(t_{(i)})} \cdot n!\prod_{i=1}^{n} h(t_{(i)})\mathrm{d}t_{(1)}\cdots\mathrm{d}t_{(n)}$$

$$= \frac{1}{n!}E\left[\frac{\prod_{i=1}^{n} f(V_{(r_i^+)}|d_i)}{\prod_{i=1}^{n} h(V_{(i)})}\right].$$

从而有定理之结果. 证毕.

§4.3 局部最强秩检验

定义 3.1 对形式参数 $\xi=(\theta,F)$，检验假设

$$H_0: \theta = 0, \qquad H_1: \theta > 0.$$

在 $F(x)$是一固定的连续分布时，若一个秩检验对每一自然的 α 水平，存在一个 $\varepsilon>0$，使得当 $0<\theta<\varepsilon$ 时，此检验是一致最强的秩检验，则称此检验为**局部最强秩检验**.

对二样本问题：设随机变量 $X_1,\cdots,X_m$ 相互独立同分布，分布函数 $F(x)$连续；随机变量 $Y_1,\cdots,Y_n$ 相互独立同分布，分布函数为 $F(x-\Delta)$. 检验假设

$$H_0: \Delta = 0, \qquad H_1: \Delta > 0.$$

在 H_0 下，$X_1,\cdots,X_m,Y_1,\cdots,Y_n$ 混合样本的秩向量 R 在集合 $\mathscr{R}=\{r|r$ 为$(1,\cdots,N)$的一个排列$\}$上均匀分布，即对任一 $r\in\mathscr{R}$，

$$P_{H_0}\{R=r\} = \frac{1}{N!} \ (P_{H_0}\{\cdot\} 表示 H_0 为真时的概率).$$

因而一个对 $\Delta>0$ 的最强的秩检验的否定域所包含的点 r，应是使 $P_\Delta\{R=r\}$（$P_\Delta\{\cdot\}$表示 Δ 为真值时的概率）取较大值的那些点，也就是从取值最大的开始，然后第二大的，…，直至取满 l 个，使 $\frac{l}{N!}=\alpha$，满足给定的水平．对任一 $r\in\mathscr{R}$，$P_\Delta\{R=r\}$是 Δ 的函数，记

$$K_r(\Delta)=P_\Delta\{R=r\}.$$

我们假定：存在 $\delta>0$，使对任一 $r\in\mathscr{R}$，在$(-\delta,\delta)$区间上一阶导数 $K_r'(\Delta)$存在且连续．那么由中值定理，有

$$K_r(\Delta)=K_r(0)+\Delta K_r'(\Delta^*)=\frac{1}{N!}+\Delta K_r'(\Delta^*),$$

其中 $0<\Delta^*<\Delta<\delta$．如此，对 $r_1,r_2\in\mathscr{R}$，有

$$\frac{1}{\Delta}[K_{r_1}(\Delta)-K_{r_2}(\Delta)]=K_{r_1}'(\Delta_1^*)-K_{r_2}'(\Delta_2^*),$$

其中 $0<\Delta_1^*<\Delta$，$0<\Delta_2^*<\Delta$，则

$$\lim_{\Delta\to0}\frac{1}{\Delta}[K_{r_1}(\Delta)-K_{r_2}(\Delta)]=K_{r_1}'(0)-K_{r_2}'(0).$$

那么当 $K_{r_1}'(0)-K_{r_2}'(0)>0$ 时，存在 $\varepsilon>0$，使 $\Delta<\varepsilon$ 时，有 $K_{r_1}(\Delta)-K_{r_2}(\Delta)>0$，因而局部最强秩检验的否定域应包含使 $K_r'(0)$取较大值的那些 r．

定理 3.1 设随机变量 $X_1,\cdots,X_m$ 相互独立同分布，分布函数 $F(x)$连续；随机变量 $Y_1,\cdots,Y_n$ 相互独立同分布，分布函数为 $F(x-\Delta)$，$F(x)$的密度函数为 $f(x)$，$f(x)$几乎处处可微，导数 $f'(x)$除可数点外连续（$f'(x)$不存在处适当补值）．又对某一$\delta>0$，存在函数 $g_1(x)$，$g_2(x)$，使当 $\Delta\in(-\delta,\delta)$和一切 x，有

$$|f'(x-\Delta)|\leqslant g_1(x),\quad f(x-\Delta)\leqslant g_2(x).$$

而

$$E_F\left[\frac{g_1(W_1)\prod_{j=2}^{n}g_2(W_j)}{\prod_{i=1}^{n}f(W_i)}\right]<\infty,\qquad(4.3.1)$$

其中随机变量 $W_1,\cdots,W_n$ 相互独立同分布，分布函数为 $F(x)$，$E_F[\cdot]$表示分布函数为 $F(x)$时的期望，且当 $\Delta>0$ 时，$f(x)$不为零的区域包含 $f(x-\Delta)$不为零的区域. 则对假设检验

$$H_0:\Delta=0,\qquad H_1:\Delta>0$$

的局部最强秩检验统计量为线性秩统计量

$$S=\sum_{i=1}^{n}a^*(R_i),\tag{4.3.2}$$

其中

$$a^*(i)=E\left[\frac{-f'(F^{-1}(U_{(i)}))}{f(F^{-1}(U_{(i)}))}\right],\quad i=1,\cdots,N,$$

$U_{(1)}<\cdots<U_{(N)}$是$[0,1]$上均匀分布的简单随机样本的次序统计量.

证明 由上一节中定理 2.3，取 $h(x)=f(x)$，则得

$$P_\Delta\{R=r\}=\frac{1}{N!}E\left[\frac{\prod_{j=1}^{n}f(V_{(r_j)}-\Delta)}{\prod_{j=1}^{n}f(V_{(r_j)})}\right],$$

其中 $V_{(1)}<\cdots<V_{(N)}$是 $f(x)$的简单随机样本的次序统计量. 注意到

$$E\left[\frac{g_1(V_1)\prod_{j=2}^{n}g_2(V_j)}{\prod_{i=1}^{n}f(V_i)}\right]=\frac{1}{N!}\sum_{r\in\mathscr{R}}E\left[\frac{g_1(V_{(r_1)})\prod_{j=2}^{n}g_2(V_{(r_j)})}{\prod_{i=1}^{n}f(V_{(r_i)})}\right],$$

故定理的条件蕴含对任一 $r\in\mathscr{R}$，有

$$E\left[\frac{g_1(V_{(r_1)})\prod_{j=2}^{n}g_2(V_{(r_j)})}{\prod_{i=1}^{n}f(V_{(r_i)})}\right]<\infty.$$

加上定理条件：$|f'(x-\Delta)|\leqslant g_1(x)$，$f(x-\Delta)\leqslant g_2(x)$，保证了可积分号下求微商. 如此

$$\frac{\partial}{\partial \Delta} P_{\Delta}\{R = r\}$$

$$= \frac{1}{N!} E\left[\frac{\sum_{k=1}^{n} \left\{ \prod_{j \neq k} f(V_{(r_j)} - \Delta) \cdot [-f'(V_{(r_k)} - \Delta)] \right\}}{\prod_{i=1}^{n} f(V_{(r_i)})} \right],$$

则在 $\Delta=0$ 点的值为

$$\frac{\partial}{\partial \Delta} P_{\Delta}\{R = r\} \Big|_{\Delta=0} = \frac{1}{N!} \sum_{k=1}^{n} E\left[\frac{-f'(V_{(r_k)})}{f(V_{(r_k)})} \right]$$

$$= \frac{1}{N!} \sum_{k=1}^{n} E\left[\frac{-f'[F^{-1}(U_{(r_k)})]}{f[F^{-1}(U_{(r_k)})]} \right] = \left[\frac{1}{N!} S \right]_{R=r}.$$

所以对应 S 取较大值的 r 组成的否定域是最强检验. 证毕.

例 1 考虑正态分布,密度函数

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < +\infty,$$

此时
$$\frac{-f'(x)}{f(x)} = -\frac{d}{dx}[\ln f(x)] = x.$$

定理 3.1 中的 $a^*(i)$ 此时为

$$a^*(i) = E\left[\frac{-f'[F^{-1}(U_{(i)})]}{f[F^{-1}(U_{(i)})]} \right] = E[F^{-1}(U_{(i)})],$$

上式中 $F(x)$ 为正态分布的分布函数,$a^*(i)$ 为正态分布样本(样本量为 N)的第 i 个次序统计量的期望值,称为正态期望值分值. 局部最强秩检验统计量为线性秩统计量

$$S = \sum_{i=1}^{n} a^*(R_i) = \sum_{i=1}^{n} E[F^{-1}(U_{(R_i)})]. \qquad (4.3.3)$$

例 2 考虑 Logistic 分布,分布函数和密度函数分别为

$$F(x) = \frac{1}{1 + e^{-x}}, \quad f(x) = \frac{e^{-x}}{(1 + e^{-x})^2}, \quad -\infty < x < \infty.$$

此时

$$\frac{-f'(x)}{f(x)} = -\frac{d}{dx}[\ln f(x)] = -\frac{d}{dx}[-x - 2\ln(1 + e^{-x})]$$

$$= 1 + \frac{2}{1+e^{-x}} \cdot (-e^{-x}) = \frac{2}{1+e^{-x}} - 1$$

$$= 2F(x) - 1.$$

定理 3.1 中的 $a^*(i)$为

$$a^*(i) = E\left[\frac{-f'[F^{-1}(U_{(i)})]}{f[F^{-1}(U_{(i)})]}\right] = 2E\{F[F^{-1}(U_{(i)})]\} - 1$$

$$= 2E\{U_{(i)}\} - 1 = \frac{2i}{N+1} - 1.$$

局部最强秩检验统计量为

$$S = \sum_{i=1}^{n} a^*(R_i) = \frac{2}{N+1}\sum_{i=1}^{n} R_i - n, \tag{4.3.4}$$

它与 $W = \sum_{i=1}^{n} R_i$ 等价. 所以 Wilcoxon 秩和检验 W 对 Logistic 分布是局部最强秩检验.

函数

$$\phi(u,f) = \frac{-f'(F^{-1}(u))}{f(F^{-1}(u))} \tag{4.3.5}$$

通常称为对应 $F(x)$的**最优分值函数**.

定理 3.2 设分布函数 $F(x)$与 $G(x)$间有关系: $F(x) = G\left(\frac{x-\mu}{\sigma}\right)$ $(-\infty<\mu<\infty, \sigma>0)$, 则对应的最优分值函数间有关系

$$\phi(u,f) = \frac{1}{\sigma}\phi(u,g),$$

其中 $f(x)$为 $F(x)$的密度函数, $g(x)$为 $G(x)$的密度函数.

证明 由 $F(x) = G\left(\frac{x-\mu}{\sigma}\right) = u$, 得

$$x = F^{-1}(u), \quad \frac{x-\mu}{\sigma} = G^{-1}(u),$$

故有

$$F^{-1}(u) = \sigma G^{-1}(u) + \mu.$$

而

$$f(x) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right), \quad f'(x) = \frac{1}{\sigma^2} g'\left(\frac{x-\mu}{\sigma}\right),$$

从而 $\phi(u,f)=\dfrac{-f'(F^{-1}(u))}{f(F^{-1}(u))}=\dfrac{-\dfrac{1}{\sigma^2}g'\left(\dfrac{F^{-1}(u)-\mu}{\sigma}\right)}{\dfrac{1}{\sigma}g\left(\dfrac{F^{-1}(u)-\mu}{\sigma}\right)}$

$$=\frac{1}{\sigma}\left(-\frac{g'(G^{-1}(u))}{g(G^{-1}(u))}\right)=\frac{1}{\sigma}\phi(u,g).$$

证毕.

定理 3.2 说明，对 $F(x)$ 作变量线性变换不影响局部最强秩检验统计量.

对一样本位置问题：设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布函数为 $F(x-\theta)$，函数 $F(t)$ 关于 0 点对称，有密度 $f(t)=f(-t)$ $(-\infty<t<\infty)$，检验假设 H_0：$\theta=0$，H_1：$\theta>0$. 一样本位置问题与二样本位置问题有类似的结果. 当 H_0 成立时，$X_1,\cdots,X_n$ 相互独立同分布，分布连续且关于 0 点对称，由第一章定理 4.2 知，计数统计量 $\Psi_1=\psi(X_1),\cdots,\Psi_n=\psi(X_n)$ 及绝对秩向量 $R^+=(R_1^+,\cdots,R_n^+)$ 相互独立，Ψ_i $(i=1,\cdots,n)$ 有参数为 $\dfrac{1}{2}$ 的两点分布，R^+ 在 $\mathscr{R}=\{(1,\cdots,n)$ 的排列$\}$ 上均匀分布. 故对给定的 $d=(d_1,\cdots,d_n)$，$d_i=0$ 或 1，$i=1,\cdots,n$，及 $r^+=(r_1^+,\cdots,r_n^+)$ 为 $(1,\cdots,n)$ 的一个排列. 事件 $\{\Psi_1=d_1,\cdots,\Psi_n=d_n,R^+=r^+\}$ 对任一 (d,r^+) 有相等的概率 $1/(2^n\cdot n!)$. 与二样本问题一样，为使功效函数值大，否定域应包含那样的 (d,r^+) 点，使在 H_1 下的概率 $K_{(d,r^+)}(\theta)=P_\theta\{\Psi_1=d_1,\cdots,\Psi_n=d_n,R^+=r^+\}$ 尽量大. 与前面二样本问题类似，在一定条件下，在 $\theta=0$ 的局部区域内找使 $K'_{(d,r^+)}(0)$ 大的 (d,r^+) 点即成. 有定理如下：

定理 3.3 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布函数为 $G(x)$，$G(x)=F(x-\theta)$，函数 $F(t)$ 关于 0 点对称，密度 $f(t)=f(-t)$ $(-\infty<t<\infty)$，且 $f(t)>0$ $(-\infty<t<\infty)$. 导数 $f'(t)$ 存在，在点 $t\neq0$ 处连续，$f'(0)=0$. $-f'(t)/f(t)$ 是 t 的非降函数. 又有 $g_1(t)$ 和 $g_2(t)$，使当 $-\delta<\theta<\delta$ 时，对一切 t 有

$$|f'(t-\theta)| \leqslant g_1(t), \quad f(t-\theta) \leqslant g_2(t),$$

且

$$E_F\left[\frac{g_1(W_1)\prod_{j=2}^{n} g_2(W_j)}{\prod_{j=1}^{n} f(W_j)}\right] < \infty, \tag{4.3.6}$$

其中随机变量 $W_1,\cdots,W_n$ 独立同分布，密度为

$$f(x|1) = \begin{cases} 0, & x \leqslant 0, \\ \dfrac{f(x)}{1-F(0)} = 2f(x), & x > 0, \end{cases}$$

则对假设检验问题 H_0: $\theta=0$, H_1: $\theta>0$ 的局部最强线性符号秩检验统计量为

$$S^+ = \sum_{i=1}^{n} \Psi_i a^*(R_i^+), \tag{4.3.7}$$

其中
$$a^*(i) = E\left[-\frac{f'(|X|_{(i)})}{f(|X|_{(i)})}\right],$$

$|X|_{(1)} < \cdots < |X|_{(n)}$ 是 $F(x)$ 的样本绝对值的次序统计量. 检验的否定域为 S^+ 大时否定 H_0.

证明 由本章定理 2.4，取其中的

$$h(x) = f(x|1),$$

则在 H_1 下的概率

$$\begin{aligned} K_{(d,r^+)}(\theta) &= P_\theta\{\Psi_1 = d_1,\cdots,\Psi_n = d_n, R^+ = r^+\} \\ &= \frac{1}{n!}\cdot\frac{1}{2^n}E\left[\frac{\prod_{i=1}^{n} f(V_{(r_i^+)} - \theta|d_i)}{\prod_{i=1}^{n} f(V_{(i)})}\right] \\ &\quad \cdot [G(0)]^{n-\sum d_i} \cdot [1-G(0)]^{\sum d_i}. \end{aligned}$$

注意到
$$G(0) = F(0-\theta) = F(-\theta);$$

$$f(x-\theta|1) = \begin{cases} 0, & x \leqslant 0, \\ \dfrac{f(x-\theta)}{1-G(0)}, & x > 0; \end{cases}$$

$$f(x-\theta|0)=\begin{cases}0, & x\leqslant 0,\\ \dfrac{f(-x-\theta)}{G(0)}=\dfrac{f(x+\theta)}{G(0)}, & x>0,\end{cases}$$

则有 $$K_{(d,r^+)}(\theta)=\frac{1}{2^n n!}E\left[\frac{\prod_{i=1}^{n}f(V_{(r_i^+)}+(-1)^{d_i}\theta)}{\prod_{i=1}^{n}f(V_{(i)})}\right].$$

定理的条件保证了可以积分号下取微商，所以

$$K'_{(d,r^+)}(\theta)=\frac{1}{2^n n!}$$

$$\times E\left[\frac{\sum_{k=1}^{n}\left\{\sum_{j\neq k}f(V_{(r_j^+)}+(-1)^{d_j}\theta)\cdot(-1)^{d_k}f'(V_{(r_k^+)}+(-1)^{d_k}\theta)\right.}{\prod_{i=1}^{n}f(V_{(i)})}\right],$$

$$\begin{aligned}K'_{(d,r^+)}(0)&=\frac{1}{2^n n!}\sum_{k=1}^{n}(-1)^{d_k}E\left[\frac{f'(V_{(r_k^+)})}{f(V_{(r_k^+)})}\right]\\&=\frac{1}{2^n n!}\left\{\sum_{k=1}^{n}\left(E\left[\frac{f'(V_{(r_k^+)})}{f(V_{(r_k^+)})}\right]-2d_kE\left[\frac{f'(V_{(r_k^+)})}{f(V_{(r_k^+)})}\right]\right)\right\}\\&=\frac{1}{2^n n!}\left\{\sum_{k=1}^{n}E\left[\frac{f'(V_{(i)})}{f(V_{(i)})}\right]+2\sum_{k=1}^{n}d_kE\left[\frac{-f'(V_{(r_k^+)})}{f(V_{(r_k^+)})}\right]\right\}.\end{aligned}$$

而 $\sum_{k=1}^{n}E\left[\frac{f'(V_{(i)})}{f(V_{(i)})}\right]$ 与 (d,r^+) 无关，所以应选 (d,r^+) 使

$$\sum_{k=1}^{n}d_kE\left[\frac{-f'(V_{(r_k^+)})}{f(V_{(r_k^+)})}\right]=S^+\Big|_{(\Psi,R^+)=(d,r^+)}.$$

取较大值的点组成否定域，这样的检验是局部最强的. 证毕.

由于 $F(x)$ 有密度且关于 0 点对称时，其样本 $X_1,\cdots,X_n$ 的绝对值 $|X_1|,\cdots,|X_n|$ 的次序统计量 $|X|_{(1)}<\cdots<|X|_{(n)}$ 有

$$|X|_{(i)}\overset{d}{=\!=}F^{-1}\left(\frac{1}{2}+\frac{1}{2}U_{(i)}\right),$$

其中 $U_{(1)}<\cdots<U_{(n)}$ 是 $(0,1)$ 均匀分布样本的次序统计量. 所以上

述定理中的 $a^*(i)$也可以表示成

$$a^*(i) = E\left[-\frac{f'\left(F^{-1}\left(\frac{1}{2}+\frac{1}{2}U_{(i)}\right)\right)}{f\left(F^{-1}\left(\frac{1}{2}+\frac{1}{2}U_{(i)}\right)\right)}\right].$$

例 3 考虑 Logistic 分布，分布函数和密度函数分别为

$$F(x) = \frac{1}{1+\mathrm{e}^{-x}},\quad f(x) = \frac{\mathrm{e}^{-x}}{(1+\mathrm{e}^{-x})^2},\quad -\infty < x < \infty.$$

此时

$$\frac{-f'(x)}{f(x)} = 2F(x) - 1,$$

定理 3.3 中的 $a^*(i)$为

$$a^*(i) = E\left[-\frac{f'\left(F^{-1}\left(\frac{1}{2}+\frac{1}{2}U_{(i)}\right)\right)}{f\left(F^{-1}\left(\frac{1}{2}+\frac{1}{2}U_{(i)}\right)\right)}\right]$$

$$= E\left[2F\left(F^{-1}\left(\frac{1}{2}+\frac{1}{2}U_{(i)}\right)\right)-1\right] = EU_{(i)} = \frac{i}{n+1},$$

它与 $a(i)=i$ 等价. 所以

$$S^+ = \sum_{i=1}^{n}\Psi_i a(R_i^+) = \sum_{i=1}^{n}\Psi_i R_i^+.$$

这就是常用的 Wilcoxon 符号秩统计量 W^+.

§4.4 功效函数的统计模拟

对假设检验问题 H_0: $\xi\in\omega$, H_1: $\xi\in\Omega\backslash\omega$,检验统计量 S，否定域 $S\geqslant c$ (c 为临界值). 其功效函数

$$\mathscr{P}_S(\xi) = P_\xi\{S\geqslant c\}$$

在非参数假设检验问题，不易求得 $\mathscr{P}_S(\xi)$的明确的表达式. 但在计算机很普及的今天，可以选择若干有代表性的 ξ 点，用统计模拟的技术计算出 $\mathscr{P}_S(\xi)$在这些点的近似值. 通过这些近似值可以比

较不同的检验方法，在一定的样本量下，看它们的功效函数的值是否有明显的差异．下面我们以一样本检验问题为例，加以说明．

对一样本位置问题：设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布为 $F(x-\theta)$，函数 $F(t)$连续关于 0 点对称，检验假设 H_0：$\theta=0$；H_1：$\theta>0$. 可有下列一些检验办法：

(1) **符号检验**（参见第一章第 2 节） 统计量

$$B=\sum_{i=1}^{n}\psi(X_i),$$

否定域

$$B\geqslant b(\alpha,n).$$

(2) **符号秩检验**（参见第一章第 4 节） 统计量

$$W^+=\sum_{i=1}^{n}\psi(X_i)R_i^+,$$

否定域

$$W^+\geqslant\omega(\alpha,n).$$

(3) ***T* 检验** 统计量

$$T=\sqrt{n}\,\overline{X}/S,\ \overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\ S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2,$$

否定域

$$T\geqslant t_{n-1}(\alpha).$$

可以用统计模拟的方法比较它们的功效，即选择典型的点(θ,F)，看这些点上功效函数的值．此例选择一些典型的常用的对称分布 $F(x)$. 例如下面一些峰度值由小到大的分布．

(1) 均匀分布．密度

$$f(t)=1,\quad \theta-\frac{1}{2}<t<\theta+\frac{1}{2},$$

期望 $\mu=\theta$，方差 $\sigma^2=\frac{1}{12}$，偏度 $s_\omega=\frac{E(X-E(X))^3}{[E(X-E(X))^2]^{3/2}}=0$，峰度 $k_u=\frac{E(X-E(X))^4}{[E(X-E(X))^2]^2}=1.8$.

(2) 正态分布．密度

$$f(t)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(t-\theta)^2},\quad -\infty<t<\infty,$$

期望 $\mu=\theta$，方差 $\sigma^2=1$，偏度 $s_\omega=0$，峰度 $k_u=3$.

(3) Logistic 分布. 密度

$$f(t) = \frac{e^{-(t-\theta)}}{[1 + e^{-(t-\theta)}]^2}, \quad -\infty < t < \infty,$$

期望 $\mu=\theta$,方差 $\sigma^2=\frac{\pi^2}{3}$,偏度 $s_w=0$,峰度 $k_u=4.2$.

(4) 双指数分布. 密度

$$f(t) = \frac{1}{2}e^{-|t-\theta|}, \quad -\infty < t < \infty,$$

期望 $\mu=\theta$,方差 $\sigma^2=2$,偏度 $s_w=0$,峰度 $k_u=6$.

(5) Cauchy 分布. 密度

$$f(t) = \frac{1}{\pi[1 + (t-\theta)^2]}, \quad -\infty < t < \infty,$$

此分布的各阶矩均不存在. 度量此分布的离散程度常取量 σ^2,使 $P\{-\sigma<X<\sigma\}=0.6826$(正态分布正负一个标准差间的概率值).

取定 $F(x)$后,θ 的值可选 $\theta=0, 0.1\sigma, 0.2\sigma, 0.3\sigma, \cdots$. 这样对给定的$(\theta,F)$点,由统计模拟方法产生分布为$(\theta,F)$的随机变量,通过计算机得到这个随机变量的一组样本值 $x_1,\cdots,x_n$,计算出检验统计量的值,即可判定这组样本值是否落入否定域. 重复产生分布(θ,F)的样本值一千次(或更多),即可得到总体为(θ,F)时,事件$\{B\geqslant b(\alpha,n)\}$,$\{W^+\geqslant\omega(\alpha,n)\}$,$\{T\geqslant t_{n-1}(\alpha)\}$在这一千次试验中出现的频率,这就是各个检验在点$(\theta,F)$处的功效函数的近似值. 可以比较值的大小,值大的功效大.

对上述一样本位置问题的三种检验办法,当样本量 n 在 10 到 30 之间时,在上述一些典型的(θ,F)点统计模拟的结果可得到下述一些结论:

总体为均匀分布: T 比较好,B 较差.

总体为正态分布: T 比 W^+稍好,两者相差不大,B 较差.

总体为 Logistic 分布: 三者接近,W^+最好.

总体为双指数分布: W^+最好.

总体为 Cauchy 分布: T 很差,B 最好.

如果在实际工作中，我们对 H_1 下的分布有一定的先验知识，如知道分布的峰度值的大小，则可选择适宜的检验方法. 有人根据样本的次序统计量 $X_{(1)},\cdots,X_{(n)}$的值，计算

$$Q_3=\frac{\overline{U}_{0.05}-\overline{M}_{0.5}}{\overline{M}_{0.5}-\overline{L}_{0.05}},\quad Q_4=\frac{\overline{U}_{0.05}-\overline{L}_{0.05}}{\overline{U}_{0.5}-\overline{L}_{0.5}}$$

来观察总体的偏度和峰度，从而决定采用何种检验办法，其中$\overline{U}_\alpha$，$\overline{M}_\alpha$，$\overline{L}_\alpha$ 分别表示 $X_{(1)},\cdots,X_{(n)}$中最大、中间、最小的 αn 个次序统计量值的平均值.

类似地用统计模拟方法，可比较二样本位置假设检验问题的两种检验方法. 假设检验问题：设随机变量 $X_1,\cdots,X_m$ 相互独立同分布，分布为 $F(x)$；随机变量 $Y_1,\cdots,Y_n$ 相互独立同分布，分布为 $F(x-\Delta)$，检验假设 H_0：$\Delta=0$，H_1：$\Delta>0$. 第一种方法采用检验统计量

$$T=\frac{\overline{Y}-\overline{X}}{\sqrt{\frac{(m-1)S_X^2+(n-1)S_Y^2}{m+n-2}\left(\frac{1}{m}+\frac{1}{n}\right)}},$$

其中$\overline{X}=\frac{1}{m}\sum_{i=1}^{m}X_i$，$\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$，$S_X^2=\frac{1}{m-1}\sum_{i=1}^{m}(X_i-\overline{X})^2$，$S_Y^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2$. 第二种方法是采用检验统计量

$$W=\sum_{i=1}^{n}R_i,$$

其中$(R_1,\cdots,R_n)$是$(Y_1,\cdots,Y_n)$在混合样本$(X_1,\cdots,X_m,Y_1,\cdots,Y_n)$中的对应的秩. 比较两种方法检验的功效，这时 $F(x)$除前述的典型的一些对称分布外，还应采用一些常见的非对称分布. 模拟的结果如下：

总体为均匀、正态分布：T 稍好.

总体为 Logistic 分布：T 与 W 差不多，W 稍好.

总体为双指数、Cauchy 分布：W 好.

总体为 Weibull 分布(形状参数为 2): T 与 W 差不多.

总体为指数分布: W 好.

上述各分布中,Weibull 分布的密度为

$$f(t) = 2te^{-t^2}, \quad t > 0,$$

期望 $\mu = \frac{\sqrt{\pi}}{2}$,方差 $\sigma^2 = 1 - \frac{\pi}{4}$,偏度 $s_u = \frac{\sqrt{\pi}}{4}(\pi - 3)\Big/\left(1-\frac{\pi}{4}\right)^{3/2}$,峰度 $k_u=\left(2-\frac{3}{8}\pi\sqrt{\pi}\right)\Big/\left(1-\frac{\pi}{4}\right)^2$.

指数分布的密度为

$$f(t) = e^{-t}, \quad t > 0,$$

期望 $\mu=1$,方差 $\sigma^2=1$,偏度 $s_u=2$,峰度 $k_u=9$.

习　题

1. 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布,分布为 $F\left(\frac{x}{\sigma}\right)$,其中分布函数 $F(t)$连续且关于 0 点对称,标准差为 1,对假设检验

$$H_0: \sigma = 1, \qquad H_1: \sigma > 1$$

有检验统计量 $S(X_1,\cdots,X_n)$,当 $S(X_1,\cdots,X_n)\geqslant c$ 时拒绝 H_0,$S(\cdot)$满足: 对任一个 $d\geqslant 1$,及任意的 $x_1,\cdots,x_n$,有

$$S(dx_1,\cdots,dx_n) \geqslant S(x_1,\cdots,x_n).$$

证明对任一 $F(t)$,此检验有对 σ 单调的功效函数.

2. 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布,密度为 $f(x-\theta)$,对一切 t,函数 $f(t)=f(-t)$. 对一样本问题

$$H_0: \theta = 0, \qquad H_1: \theta > 0$$

可以用一个线性符号秩统计量 $S^+=\sum_{i=1}^{n}\Psi_i a(R_i^+)$作检验. 说明基于 S^+的检验的功效函数对于 θ 单调.

3. 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布,分布为 $F(x-\theta)$,函数 $F(t)$连续且关于 0 点对称. 对一样本问题

$$H_0: \theta = 0, \qquad H_1: \theta > 0$$

可以用 Wilcoxon 符号秩统计量构成检验. 说明此检验对一切 $P\{X_1+X_2>0\}>\frac{1}{2}$ 的备择假设是相合的.

4. 对双侧否定域$\{T_N\geqslant c_N$ 或 $T_N\leqslant c'_N$ 时否定 $H_0\}$,叙述定理 1.3 类似的定理,并证明之.

5. 证明下列定理:设随机变量 $X_1,\cdots,X_n$ 相互独立同分布,分布函数 $F(x)$连续. 假设检验问题

$$H_0:F(0)=\frac{1}{2},\qquad H_1:F(x)\text{ 的中位数大于 }0$$

有检验统计量 $S(X_1,\cdots,X_n)$,否定域为 $S(\cdot)\geqslant c$,又对任意的 $x_i\leqslant x'_i\ (i=1,\cdots,n)$, $S(\cdot)$满足

$$S(x'_1,\cdots,x'_n)\geqslant S(x_1,\cdots,x_n),$$

则对任意 $F(x)$,$G(x)$,当 $F(x)$随机地小于 $G(x)$时,有

$$\mathscr{P}_S(F)\leqslant\mathscr{P}_S(G).$$

6. 对二样本位置参数问题,若分布函数 $F(x)$有密度

$$f(x)=\exp\{x-\mathrm{e}^x\},\quad -\infty<x<+\infty.$$

试找出线性秩统计量最优分值函数.

7. 设随机变量 $X_1,\cdots,X_m$ 相互独立同分布,分布为 $F(x)$;随机变量 $Y_1,\cdots,Y_n$ 相互独立同分布,分布为 $F\left(\frac{x}{\eta}\right)$,分布函数 $F(t)$ 连续,且 $F(0)=\frac{1}{2}$,对二样本尺度参数假设检验问题

$$H_0:\eta=1,\qquad H_1:\eta>1$$

的最优分值函数经分布变量尺度变换保持不变,即若 $G(x/\sigma)=F(x)$,则对 $G(x)$的二样本尺度参数问题的线性秩统计量最优分值函数与对 $F(x)$的一样.

第五章　检验的渐近相对效率

§5.1　Pitman 渐近相对效率

对假设检验问题

$$H_0: \theta \in \omega, \qquad H_1: \theta \in \Omega \backslash \omega.$$

若有$\{S_n\}$，$\{T_n\}$两列检验统计量，其中 n 为统计量使用的样本量.设$\{S_n\}$，$\{T_n\}$显著水平为 α 的否定域分别为 C_n 和 D_n，即有

$$P_\theta\{S_n \in C_n\} \leqslant \alpha, \quad 对一切 \theta \in \omega;$$

$$P_\theta\{T_n \in D_n\} \leqslant \alpha, \quad 对一切 \theta \in \omega.$$

比较此二列检验，通常比较它们的功效函数

$$\mathscr{P}_{S_n}(\theta) = P_\theta\{S_n \in C_n\}, \quad \theta \in \Omega \backslash \omega;$$

$$\mathscr{P}_{T_n}(\theta) = P_\theta\{T_n \in D_n\}, \quad \theta \in \Omega \backslash \omega.$$

对于特定的 $\theta^* \in \Omega \backslash \omega$，$\mathscr{P}_{S_n}(\theta^*)$，$\mathscr{P}_{T_n}(\theta^*)$大者为优. 一般使用的样本量越大，功效函数的值会越高. 因而也可规定功效的值，看达到此功效所需的样本量. 所需样本量少者为优. Pitman 渐近相对效率，就是在零假设的一个邻近区域内，规定功效的值，比较所需的样本量.

定义 1.1　对假设检验问题 H_0: $\theta=\theta_0$，H_1: $\theta\neq\theta_0$，取一列特定的备择假设 $\theta=\theta_i$ $(i=1,2,\cdots)$，$\theta_i\neq\theta_0$ 且 $\lim\limits_{i\to\infty}\theta_i=\theta_0$. S_{n_i}和 T_{m_i}是对应于假设检验问题

$$H_0: \theta = \theta_0, \qquad H_1: \theta = \theta_i$$

的检验统计量，n_i 和 m_i 为对应的样本量，满足

$$\lim_{i\to\infty}\mathscr{P}_{S_{n_i}}(\theta_0) = \lim_{i\to\infty}\mathscr{P}_{T_{m_i}}(\theta_0) = \alpha,$$

$$\alpha < \lim_{i\to\infty}\mathscr{P}_{S_{n_i}}(\theta_i) = \lim_{i\to\infty}\mathscr{P}_{T_{m_i}}(\theta_i) < 1.$$

若对一切满足上述条件的$\{\theta_i\}$，$\{n_i\}$，$\{m_i\}$存在相同的

$$\lim_{i\to\infty}\frac{m_i}{n_i},$$

则称此极限值为$\{S_n\}$相对$\{T_n\}$的 Pitman **渐近相对效率**，记为 $\mathrm{ARE}(S,T)$.

关于渐近相对效率的计算，下面的 Noether 定理是最基本的.

定理 1.1(Noether 定理) 对假设检验问题 H_0: $\theta=\theta_0$, H_1: $\theta=\theta_i$ $(i=1,2,\cdots)$且 $\lim_{i\to\infty}\theta_i=\theta_0$, $\{S_{n_i}\}$和$\{T_{m_i}\}$是两列检验统计量，若有与它们相联系的数列$\{\mu_{S_{n_i}}(\theta)\}$，$\{\mu_{T_{m_i}}(\theta)\}$，$\{\sigma_{S_{n_i}}(\theta)\}$，$\{\sigma_{T_{m_i}}(\theta)\}$满足下列假定$(\mathrm{A}_1)\sim(\mathrm{A}_6)$.

(A_1) $\dfrac{S_{n_i}-\mu_{S_{n_i}}(\theta_i)}{\sigma_{S_{n_i}}(\theta_i)}$ 和 $\dfrac{T_{m_i}-\mu_{T_{m_i}}(\theta_i)}{\sigma_{T_{m_i}}(\theta_i)}$当 $\theta=\theta_i$ 为真时，有相同的取值范围及相同的连续极限分布 $H(\cdot)$.

(A_2) 当$(\mathrm{A})_1$ 中 θ_i 代之为 θ_0 时，有同样假定.

(A_3) $\lim_{i\to\infty}\dfrac{\sigma_{S_{n_i}}(\theta_i)}{\sigma_{S_{n_i}}(\theta_0)}=\lim_{i\to\infty}\dfrac{\sigma_{T_{m_i}}(\theta_i)}{\sigma_{T_{m_i}}(\theta_0)}=1.$

(A_4) $\dfrac{\mathrm{d}}{\mathrm{d}\theta}\mu_{S_{n_i}}(\theta)$和$\dfrac{\mathrm{d}}{\mathrm{d}\theta}\mu_{T_{m_i}}(\theta)$存在，且在 $\theta=\theta_0$ 的某一闭域内连续. 导数 $\mu'_{S_{n_i}}(\theta_0)$，$\mu'_{T_{m_i}}(\theta_0)$不为 0.

(A_5) $\lim_{i\to\infty}\dfrac{\mu'_{S_{n_i}}(\theta_i)}{\mu'_{S_{n_i}}(\theta_0)}=\lim_{i\to\infty}\dfrac{\mu'_{T_{m_i}}(\theta_i)}{\mu'_{T_{m_i}}(\theta_0)}=1.$

(A_6) $\lim_{n\to\infty}\dfrac{\mu'_{S_n}(\theta_0)}{\sqrt{n\sigma^2_{S_n}(\theta_0)}}=K_S$, $\lim_{n\to\infty}\dfrac{\mu'_{T_n}(\theta_0)}{\sqrt{n\sigma^2_{T_n}(\theta_0)}}=K_T$,

其中 K_S 和 K_T 为正的常数，则 Pitman 渐近相对效率

$$\mathrm{ARE}(S,T) = [K_S/K_T]^2. \tag{5.1.1}$$

证明 设$\{S_{n_i}\}$和$\{T_{m_i}\}$对应的否定域分别为$\{S_{n_i}\geqslant c_{n_i}\}$和$\{T_{m_i}$

$\geqslant d_{m_i}\}$，满足渐近相对效率的要求，即有

$$\lim_{i\to\infty}\mathscr{P}_{S_{n_i}}(\theta_0)=\lim_{i\to\infty}\mathscr{P}_{T_{m_i}}(\theta_0)=\alpha,$$

$$\lim_{i\to\infty}\mathscr{P}_{S_{n_i}}(\theta_i)=\lim_{i\to\infty}\mathscr{P}_{T_{m_i}}(\theta_i)=\beta,$$

则由(A_1)，(A_2)得

$$\lim_{i\to\infty}\frac{c_{n_i}-\mu_{S_{n_i}}(\theta_i)}{\sigma_{S_{n_i}}(\theta_i)}=\lim_{i\to\infty}\frac{d_{m_i}-\mu_{T_{m_i}}(\theta_i)}{\sigma_{T_{m_i}}(\theta_i)}=H^{-1}(1-\beta),$$

$$\lim_{i\to\infty}\frac{c_{n_i}-\mu_{S_{n_i}}(\theta_0)}{\sigma_{S_{n_i}}(\theta_0)}=\lim_{i\to\infty}\frac{d_{m_i}-\mu_{T_{m_i}}(\theta_0)}{\sigma_{T_{m_i}}(\theta_0)}=H^{-1}(1-\alpha).$$

又由(A_3)可得

$$\lim_{i\to\infty}\frac{c_{n_i}-\mu_{S_{n_i}}(\theta_i)}{\sigma_{S_{n_i}}(\theta_0)}=\lim_{i\to\infty}\frac{d_{m_i}-\mu_{T_{m_i}}(\theta_i)}{\sigma_{T_{m_i}}(\theta_0)}=H^{-1}(1-\beta),$$

从而可得

$$\lim_{i\to\infty}\frac{\mu_{S_{n_i}}(\theta_i)-\mu_{S_{n_i}}(\theta_0)}{\sigma_{n_i}(\theta_0)}=\lim_{i\to\infty}\frac{\mu_{T_{m_i}}(\theta_i)-\mu_{T_{m_i}}(\theta_0)}{\sigma_{T_{m_i}}(\theta_0)}.$$

利用定理之条件(A_4)，将上式极限号下之分子用中值定理，并相除即得

$$\lim_{i\to\infty}\left[\frac{\mu'_{S_{n_i}}(\theta_i^*)}{\sigma_{S_{n_i}}(\theta_0)}\Big/\frac{\mu'_{T_{m_i}}(\theta_i^{**})}{\sigma_{T_{m_i}}(\theta_0)}\right]=1,$$

其中 $\theta_i^*,\theta_i^{**}\in(\theta_0,\theta_i)$ $(i=1,2,\cdots)$. 再由(A_5)，(A_6)，即得

$$\lim_{i\to\infty}\left[\sqrt{n_i}K_S\Big/\sqrt{m_i}K_T\right]=1,$$

因此

$$\mathrm{ARE}(S,T)=\lim_{i\to\infty}\frac{m_i}{n_i}=\left[\frac{K_S}{K_T}\right]^2.$$

证毕.

定义 1.2 对假设检验问题 H_0: $\theta=\theta_0$, H_1: $\theta=\theta_i$，上述定理

中的量

$$K_S = \lim_{n\to\infty} \frac{\mu'_{S_n}(\theta_0)}{\sqrt{n\sigma^2_{S_n}(\theta_0)}} \tag{5.1.2}$$

称为 S_n 的**效力因子**，记为 eff(S).

例 1 考虑总体为正态分布

$$f(x;\theta,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2}, \quad -\infty < x < \infty,$$

检验假设 H_0: $\theta=0$, H_1: $\theta=\theta_i$ ($i=1,2,\cdots$), $\lim\limits_{i\to\infty}\theta_i=0$. 考虑检验统计量

$$T = \sqrt{n}\,\overline{X}/S, \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

$$B = \sum_{i=1}^{n}\psi(X_i) \text{（符号统计量，参见第一章第 2 节），}$$

则取

$$\mu_{T_n}(\theta) = \frac{\sqrt{n}\,\theta}{\sigma}, \quad \sigma_{T_n}(\theta) = 1,$$

$$\mu_{B_n}(\theta) = np, \quad \sigma_{B_n}(\theta) = \sqrt{np(1-p)},$$

其中

$$p = \frac{1}{\sqrt{2\pi}\sigma}\int_0^{\infty} e^{-\frac{1}{2}\left(\frac{t-\theta}{\sigma}\right)^2} dt.$$

容易验证它们满足 Noether 定理的条件(A_1)~(A_6)，且

$$\mu'_{T_n}(\theta) = \frac{\sqrt{n}}{\sigma},$$

$$\mu'_{B_n}(\theta) = \frac{n}{\sqrt{2\pi}\sigma}\int_0^{\infty}\frac{1}{\sigma^2}(t-\theta)e^{-\frac{1}{2}\left(\frac{t-\theta}{\sigma}\right)^2}dt$$

$$= \frac{n}{\sqrt{2\pi}\sigma}\int_0^{\infty} d\left(-e^{-\frac{1}{2}\left(\frac{t-\theta}{\sigma}\right)^2}\right) = \frac{n}{\sqrt{2\pi}\sigma}e^{-\frac{\theta^2}{2\sigma^2}},$$

从而效力因子

$$\text{eff}(T) = \frac{1}{\sigma},$$

$$\text{eff}(B)=\lim_{n\to\infty}\frac{\mu'_{B_n}(0)}{\sqrt{n\sigma_{B_n}(0)}}=\lim_{n\to\infty}\left[\frac{n}{\sqrt{2\pi}\sigma}\Big/\frac{n}{2}\right]=\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}.$$

故渐近相对效率

$$\text{ARE}(T,B)=\left[\frac{1}{\sigma}\Big/\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right]^2=\frac{\pi}{2}.$$

Noether 定理的核心条件是在 H_0 和 H_1 下，检验统计量序列均有渐近分布. 为了应用 Noether 定理，我们需要对经常涉及的 U 统计量和线性秩统计量在 H_1 下的极限分布作进一步的讨论.

§5.2　推广的 U 统计量的极限定理

在 Pitman 备择假设 H_1: $\theta=\theta_i$ $(i=1,2,\cdots)$, $\lim\limits_{i\to\infty}\theta_i=\theta_0$ 下，总体分布随 θ_i 变动，故需考虑总体变动时 U 统计量的极限分布. 此时简单随机样本

$$X_{1,n},\cdots,X_{n,n}\sim F(x;\theta_n),$$

U 统计量

$$U_n(X_{1,n},\cdots,X_{n,n})=\frac{1}{\binom{n}{r}}\sum_{\substack{(\beta_1,\cdots,\beta_r)\\ \subset(1,\cdots,n)}}h(X_{\beta_1,n},\cdots,X_{\beta_r,n}),$$

$$E[U_n(X_{1,n},\cdots,X_{n,n})]=\theta_n.$$

记

$$\begin{aligned}h_{1,n}(x)&=E[h(X_{1,n},X_{2,n},\cdots,X_{r,n})\,|\,X_{1,n}=x]\\&=E[h(x,X_{2,n},\cdots,X_{r,n})],\end{aligned}\tag{5.2.1}$$

则有

$$\begin{aligned}\zeta_{1,n}&=\text{cov}[h(X_{1,n},X_{2,n},\cdots,X_{r,n}),h(X_{1,n},X_{r+1,n},\cdots,X_{2r-1,n})]\\&=\text{var}[h_{1,n}(X_{1,n})].\end{aligned}$$

与第二章固定总体情形下，U 统计量的推导类似，易知 $U_n(X_{1,n},\cdots,X_{n,n})-\theta_n$在变量集合

$$\mathscr{V}_n = \left\{ \sum_{i=1}^{n} K(X_{i,n}),\ K(t)\ 为实函数 \right\}$$

上的投影为

$$V_n^* = \frac{r}{n}\sum_{i=1}^{n}[h_{1,n}(X_{i,n}) - \theta_n]. \qquad (5.2.2)$$

定理 2.1 对上述 U 统计量,若存在常数 $M>0$,使对任一 $n \geqslant r$,有

$$E[h^2(X_{1,n},\cdots,X_{r,n})] \leqslant M,$$

则

$$\lim_{n\to\infty}\{n\mathrm{var}[U_n(X_{1,n},\cdots,X_{r,n})] - r^2\zeta_{1,n}\} = 0.$$

证明 由定理条件,有

$$\zeta_{r,n} = E[h^2(X_{1,n},\cdots,X_{r,n})] - \theta_n^2 \leqslant M$$

存在且有限,所以 $|\zeta_{c,n}| \leqslant \zeta_{r,n} \leqslant M$ $(c=1,\cdots,r)$ 皆存在,而由 U 统计量方差的一般表达式

$$\mathrm{var}[U_n(X_{1,n},\cdots,X_{n,n})] = \frac{1}{\binom{n}{r}}\sum_{c=1}^{r}\binom{r}{c}\binom{n-r}{r-c}\zeta_{c,n}$$

$$= \sum_{c=1}^{r}\frac{(r!)^2}{c![(r-c)!]^2}\cdot\frac{(n-r)\cdots(n-2r+c+1)}{n(n-1)\cdots(n-r+1)}\zeta_{c,n}.$$

因 $|\zeta_{c,n}| \leqslant M$ 对任一 $n \geqslant r$ 均成立,所以当 $n\to\infty$ 时,仅有 c 取值 1 的项是 $\frac{1}{n}$ 阶的,其余的项的阶均高于 $\frac{1}{n^2}$,故

$$n\mathrm{var}(U_n) - r^2\zeta_{1,n} \to 0.$$

证毕.

定理 2.2 设随机变量 $X_{1,n},\cdots,X_{n,n}$ 相互独立同分布,分布为 $F(x-\theta_n)$. 若对任一 $n \geqslant r$ 有常数 $M>0$,使

$$E[|h_{1,n}(X_{1,n}) - \theta_n|^3] \leqslant M.$$

又

$$\lim_{n\to\infty} r^2\zeta_{1,n} = \eta,$$

其中 η 是大于 0 的常数，则 $\sqrt{n}V_n^* \Big/ \sqrt{r^2\zeta_{1,n}}$ 有极限分布 $N(0,1)$.

证明 利用 Berry-Essean 中心极限定理即刻可得定理结论. 该中心极限定理说：对任一 n，若 $W_{1,n},\cdots,W_{n,n}$ 是期望为零方差为 1 的独立同分布随机变量列，如果

$$\lim_{n\to\infty} n^{-\frac{1}{2}} \cdot E[|W_{1,n}|^3] = 0,$$

则对一切 x 和 n 存在不依赖 n 的常数 A，使

$$\left| P\left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} W_{i,n} \leqslant x \right\} - \Phi(x) \right| \leqslant A n^{-\frac{1}{2}} E[|W_{1,n}|^3],$$

其中 $\Phi(x)$ 为标准正态分布函数. 我们只要将定理中的投影 V_n^* 的各加项 $r[h_{1,n}(X_{i,n})-\theta_n] \Big/ \sqrt{r^2\zeta_{1,n}}$ 视为 Berry-Essean 中心极限定理中的 $W_{i,n}$ 即可. 证毕.

定理 2.3 设随机变量 $X_{1,n},\cdots,X_{n,n}$ 相互独立同分布，分布函数为 $F(x-\theta_n)$. 若

(i) 对任一 $n \geqslant r$，有

$$E[h^2(X_{1,n},\cdots,X_{r,n})] \leqslant \infty;$$

(ii) $\lim\limits_{n\to\infty}[n\text{var}\{U_n(X_{1,n},\cdots,X_{n,n})\}] = \lim\limits_{n\to\infty} r^2\zeta_{1,n} = \eta$，其中 η 是大于 0 的常数；

(iii) 对任一 $n \geqslant r$，有常数 $M>0$，使

$$E[|h_{1,n}(X_{1,n}) - \theta_n|^3] \leqslant M,$$

则 $\sqrt{n}[U_n(X_{1,n},\cdots,X_{n,n})-\theta_n]$ 有极限分布 $N(0,1)$.

证明 与第二章定理 2.4 之证明类似. 可证 $\sqrt{n}[U_n(X_{1,n},\cdots,X_{n,n})-\theta_n]$ 与其投影 $\sqrt{n}V_n^*$ 有相同的极限分布. 再由本节定理 2.2 立即可推得本定理结论. 证毕.

例 1 上述定理可用于一样本问题：$F(x-\beta)$ 连续，关于 β 对称. 检验假设

$$H_0: \beta = 0, \qquad H_1: \beta = \beta_i = \frac{c}{\sqrt{n_i}}, \quad c > 0.$$

在 H_1 下：样本 $X_{1,n_i},\cdots,X_{n_i,n_i}$ 相互独立同分布，分布函数为 $F\left(x-\frac{c}{\sqrt{n_i}}\right)$. 例如对 Wilcoxon 符号秩检验统计量 W^+ 的主要部分的 U 统计量

$$U(X_{1,n_i},\cdots,X_{n_i,n_i})=\frac{1}{\binom{n_i}{2}}\sum_{l=1}^{n_i}\sum_{l<j}\psi(X_{l,n_i}+X_{j,n_i})$$

相应的参数 θ_n, U 统计量的核 $h(\cdot)$ 及其投影 $h_{1,n_i}(\cdot)$ 如下：

$$\theta_n=P_{\beta_i}\{X_{1,n_i}+X_{2,n_i}>0\};$$

$$0\leqslant h(X_{1,n_i},X_{2,n_i})=\psi(X_{1,n_i}+X_{2,n_i})\leqslant 1;$$

$$\begin{aligned}0\leqslant h_{1,n_i}(x)&=E\{h(X_{1,n_i},X_{2,n_i})\,|\,X_{1,n_i}=x\}\\&=E\{\psi(x+X_{2,n_i})\}=P_{\beta_i}\{x+X_{2,n_i}>0\}\\&=1-F\left(-x-\frac{c}{\sqrt{n_i}}\right)\leqslant 1,\end{aligned}$$

可知定理 2.3 的条件(i)和(iii)成立. 又

$$\begin{aligned}&\lim_{n\to\infty}[n\mathrm{var}\{U_n(X_{1,n},\cdots,X_{n,n})\}]\\&=\lim_{n\to\infty}r^2\zeta_{1,n}=2^2\lim_{n\to\infty}\mathrm{var}\{h_{1,n}(X_{1,n})\}\\&=4\lim_{n\to\infty}\mathrm{var}\left\{1-F\left(-X_{1,n}-\frac{c}{\sqrt{n}}\right)\right\}\\&=4\lim_{n\to\infty}\mathrm{var}\left\{F\left(-X_{1,n}-\frac{c}{\sqrt{n}}\right)\right\}\\&=4\lim_{n\to\infty}\mathrm{var}\left\{F\left(X_{1,n}+\frac{c}{\sqrt{n}}\right)\right\}\\&=4\cdot\mathrm{var}\{F(X)\}=4\cdot\frac{1}{12}=\frac{1}{3},\end{aligned}$$

其中随机变量 $X\sim F(x)$. 定理的条件(ii)满足，所以

$$\sqrt{n_i}\,\{U_{n_i}(X_{1,n_i},\cdots,X_{n_i,n_i})-P_{\beta_i}\{X_{1,n_i}+X_{2,n_i}>0\}\}$$

有极限分布 $N\left(0,\frac{1}{3}\right)$.

对二样本问题有类似的定理.

定理 2.4 设样本 $X_{1,m},\cdots,X_{m,m}$相互独立同分布，分布函数 $F_m(x)$连续；$Y_{1,n},\cdots,Y_{n,n}$相互独立同分布，分布函数 $G_n(x)$连续. 统计量 $U(X_{1,m},\cdots,X_{m,m};Y_{1,n},\cdots,Y_{n,n})$是自由度$(r,s)$的可估参数 $\theta_{m,n}$的 U 统计量，其对称核为 $h(X_{1,m},\cdots,X_{r,m};Y_{1,n},\cdots,Y_{s,n})$. 若

(i) $E[h^2(X_{1,m},\cdots,X_{r,m};Y_{1,n},\cdots,Y_{s,n})]<\infty$;

(ii) $\lim\limits_{N\to\infty} N\cdot\mathrm{var}[U(X_{1,m},\cdots,X_{m,m};Y_{1,n},\cdots,Y_{n,n})]$

$$=\lim_{N\to\infty}\left[\frac{r^2\zeta_{1,0,m,n}}{m/N}+\frac{s^2\zeta_{0,1,m,n}}{n/N}\right]=\eta,$$

其中 $N=m+n$, $m/N\to\lambda\ (N\to\infty)$;

(iii) $$E[|h_{1,0,m,n}(X_{1,m})-\theta_{m,n}|^3]<M_1,$$
$$E[|h_{0,1,m,n}(Y_{1,n})-\theta_{m,n}|^3]<M_2,$$

其中 M_1,M_2 是大于 0 的常数，

则 $$\sqrt{N}\{U(X_{1,m},\cdots,X_{m,m};Y_{1,n},\cdots,Y_{n,n})-\theta_{m,n}\}$$

有极限分布 $N(0,\eta)$.

例 2 将定理 2.4 用于二样本问题：样本 $X_{1,m_i},\cdots,X_{m_i,m_i}$相互独立同分布，分布函数 $F(x)$连续；$Y_{1,n_i},\cdots,Y_{n_i,n_i}$相互独立同分布，分布函数 $F(x-\Delta)$连续. 检验假设

$$H_0:\ \Delta=0,\qquad H_1:\ \Delta=\Delta_i=\frac{c}{\sqrt{N_i}},\quad N_i=m_i+n_i.$$

当$\frac{m_i}{N_i}\to\lambda$，在这样的备择假设下，可得 Wilcoxon 秩和统计量 W(参见第一章第 3 节)，有

$$\sqrt{N_i}\left\{\frac{1}{m_in_i}\sum_{l=1}^{n_i}\sum_{j=1}^{m_i}\psi(Y_{l,n_i}-X_{j,m_i})-P_{\Delta_i}\{X_{1,m_i}<Y_{1,n_i}\}\right\}$$

的极限分布为 $N\left(0,\frac{1}{12\lambda(1-\lambda)}\right)$.

例 3 ARE 的例子.

考虑一样本问题：随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布函数为$F(x-\theta)$，函数 $F(t)$连续，关于 0 点对称，检验假设

$$H_0:\theta=\theta_0,\qquad H_1:\theta>\theta_0.$$

Pitman 推移备择假设为 $\theta=\theta_i=\theta_0+\frac{c}{\sqrt{n_i}}$ $(i=1,2,\cdots)$，$n_i\to\infty$. 讨论三个检验统计量：

检验 1　统计量

$$T_n=\frac{\sqrt{n}\,(\overline{X}-\theta_0)}{S}.$$

当 $T_n\geqslant t_{\alpha,n-1}$时，否定 H_0，$t_{\alpha,n-1}$是显著水平为 α 的临界值.

检验 2　统计量

$$W_n^+=\sum_{i=1}^n\sum_{i\leqslant j}\psi(X_i+X_j-2\theta_0).$$

当 $W_n^+\geqslant\omega^+(\alpha,n)$时，否定 H_0，$\omega^+(\alpha,n)$是显著水平为 α 的临界值.

检验 3　统计量

$$B_n=\sum_{i=1}^n\psi(X_i-\theta_0).$$

当 $B_n\geqslant b\left(\alpha,n,\frac{1}{2}\right)$时，否定 H_0，$b\left(\alpha,n,\frac{1}{2}\right)$是显著水平为 α 的临界值. 用 Noether 定理计算渐近相对效率 ARE. 可先计算它们各自的效力因子 $\mathrm{eff}(T_n)$，$\mathrm{eff}(W_n^+)$，$\mathrm{eff}(B_n)$，得

$$\mathrm{eff}(T_n)=\frac{1}{\sigma},\quad \mathrm{eff}(W_n^+)=2\sqrt{3}\int_{-\infty}^{+\infty}f^2(y)\mathrm{d}y,$$

$$\mathrm{eff}(B_n)=2f(0),$$

其中 σ 为分布 $F(x)$的标准差，$f(x)$为 $F(x)$的密度函数.

作为例子，我们计算 $\mathrm{eff}(W_n^+)$. 首先要验证 Noether 定理的条件均满足. 不妨设 $\theta_0=0$. 假定 $f(x)=\frac{\mathrm{d}}{\mathrm{d}x}F(x)$存在、有界，除可数点外连续，在 $x=0$ 点连续. 此时，令

$$V_n^+=\frac{1}{\binom{n}{2}}W_n^+=\frac{1}{\binom{n}{2}}\sum_{i=1}^n\psi(X_i)+\frac{1}{\binom{n}{2}}\sum_{i=1}^n\sum_{i<j}\psi(X_i+X_j),$$

则基于 W_n^+ 的检验与基于 V_n^+ 的检验等价. 取

$$\mu_n(\theta) = P_\theta\{X_1 + X_2 > 0\}, \quad \sigma_n(\theta) = \frac{1}{\sqrt{3n}},$$

由 U 统计量的性质知，在假设 H_0: $\theta=\theta_0=0$ 或 H_1: $\theta=\theta_i=\frac{c}{\sqrt{n_i}}$ 下，均有

$$\frac{V_{n_i} - \mu_{n_i}(\theta_i)}{\sigma_{n_i}(\theta_i)}, \quad i = 0,1,2,\cdots$$

有极限分布 $N(0,1)$，即 Noether 定理的条件(A_1)，(A_2)满足. $\sigma_n(\theta)=\frac{1}{\sqrt{3n}}$与 θ 无关，所以(A_3)也满足. 关于(A_4)有

$$\begin{aligned}\mu_n(\theta) &= \int_{-\infty}^{+\infty}[1 - F(-x-\theta)]\mathrm{d}F(x-\theta) \\ &= \int_{-\infty}^{+\infty}[1 - F(-y-2\theta)]\mathrm{d}F(y),\end{aligned}$$

由于 $f(x)$有界，由控制收敛定理知可积分号下取微商

$$\frac{\mathrm{d}\mu_n(\theta)}{\mathrm{d}\theta} = \int_{-\infty}^{+\infty}2f(-y-2\theta)]\mathrm{d}F(y).$$

当 $\theta=0$ 时，$F(x)$关于 0 点对称，$f(x)=f(-x)$，所以

$$\left.\frac{\mathrm{d}\mu_n(\theta)}{\mathrm{d}\theta}\right|_{\theta=0} = 2\int_{-\infty}^{+\infty}f^2(y)\mathrm{d}y.$$

由于 $f(x)$在 $x=0$ 连续且仅有可数个不连续点，故(A_4)满足. 又

$$\begin{aligned}\left.\frac{\mathrm{d}\mu_n(\theta)}{\mathrm{d}\theta}\right|_{\theta=\frac{c}{\sqrt{n}}} &= 2\int_{-\infty}^{+\infty}f\left(-y-\frac{2c}{\sqrt{n}}\right)f(y)\mathrm{d}y \\ &= 2\int_{-\infty}^{+\infty}f\left(y+\frac{2c}{\sqrt{n}}\right)f(y)\mathrm{d}y \\ &\to 2\int_{-\infty}^{+\infty}f^2(y)\mathrm{d}y \quad (n\to\infty),\end{aligned}$$

所以(A_5)满足. 而关于(A_6)有

$$\lim_{n\to\infty}\frac{\mu_n'(0)}{\sqrt{n\sigma_n^2(0)}} = \frac{2\int_{-\infty}^{+\infty}f^2(y)\mathrm{d}y}{1/\sqrt{3}} = 2\sqrt{3}\int_{-\infty}^{+\infty}f^2(y)\mathrm{d}y,$$

从而 Noether 定理的条件均满足，因此有

$$\mathrm{eff}(W_n^+) = 2\sqrt{3}\int_{-\infty}^{+\infty} f^2(y)\mathrm{d}y.$$

由 T_n, W_n^+, B_n 的效力因子可算得它们的渐近相对效率(如下表).

分　布	ARE(W^+, T)	ARE(B, T)
均　匀	1	1/3
正　态	$3/\pi$	$2/\pi$
Logistic	$\pi^2/9$	$\pi^2/12$
双指数	3/2	2

类似地，对二样本问题，当 $f(x)$ 有界，除可数点外连续，且在 $x=0$ 连续时，可算得

$$\mathrm{eff}(W) = \sqrt{12\lambda(1-\lambda)}\int_{-\infty}^{+\infty} f^2(x)\mathrm{d}x,$$

$$\mathrm{eff}(T) = \sqrt{\lambda(1-\lambda)}\,\frac{1}{\sigma},$$

其中 $0<\lambda=\lim\limits_{N\to\infty}\frac{m}{N}<1$，$W$ 为 Wilcoxon 秩和统计量，T 为二样本均值检验 t-统计量，

$$T = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{(m-1)S_X^2+(n-1)S_Y^2}{m+n-2}\left(\frac{1}{n}+\frac{1}{m}\right)}},$$

其中 $\overline{X}, \overline{Y}, S_X^2, S_Y^2$ 分别为两个样本的样本均值和样本方差. 从而可算得它们的渐近相对效率(如下表).

分　布	ARE(W, T)
均　匀	1
正　态	$3/\pi$
Logistic	$\pi^2/9$
双指数	3/2
指　数	3

§ 5.3 二样本位置问题线性秩统计量的渐近相对效率

对二样本位置问题：设样本 $X_1,\cdots,X_m$ 相互独立同分布，分布函数 $F(x)$ 连续；$Y_1,\cdots,Y_n$ 相互独立同分布，分布为 $F(x-\Delta)$. 检验假设 H_0: $\Delta=0$, H_1: $\Delta>0$ 可以有许多线性秩统计量适于作为检验统计量. 实际上我们可以定义各式各样的分值，只要 $a(1)\leqslant a(2)\leqslant\cdots\leqslant a(N)$，而回归常数 $c(i)$ 取为

$$c(i)=\begin{cases}0, & i=1,\cdots,m,\\ 1, & i=m+1,\cdots,m+n(=N),\end{cases}$$

那么线性秩统计量

$$S=\sum_{i=1}^{n}a(R_i),$$

其中秩向量 $R=(Q_1,\cdots,Q_m,R_1,\cdots,R_n)$ 是混合样本 $(X_1,\cdots,X_m,Y_1,\cdots,Y_n)$ 对应的秩向量. 这样的线性秩统计量均可用来检验上述二样本位置问题，其否定域的形式为 $S>c$. 临界值 c 由 S 是 H_0 下的分布和显著水平 α 确定. 例如取 $a(i)=i$ $(i=1,\cdots,N)$，则得 Wilcoxon 秩和统计量

$$W=\sum_{i=1}^{n}R_i;$$

取 $a(i)$ 为

$$a_M(i)=\begin{cases}1, & i>\dfrac{N+1}{2},\\[2ex] 0, & i\leqslant\dfrac{N+1}{2},\end{cases}\tag{5.3.1}$$

则得线性秩统计量

$$\begin{aligned}M&=\sum_{i=1}^{n}a_M(R_i)\\ &=\{Y_1,\cdots,Y_n\text{ 中比混合样本中位数大的个数}\}.\end{aligned}$$

又对任一连续分布函数 $F_0(x)$，定义分值

$$a_{F_0}(i) = F_0^{-1}\left(\frac{i}{N+1}\right) \tag{5.3.2}$$

或

$$a_{F_0}^*(i) = E_{F_0}(F_0^{-1}(U_{(i)})), \tag{5.3.3}$$

其中 $F_0^{-1}(t)$是 F_0 的分位数函数. $U_{(1)} \leqslant \cdots \leqslant U_{(N)}$是(0,1)均匀分布样本的次序统计量. $a_{F_0}(i)$称为 F_0 分位数分值. $a_{F_0}^*(i)$称为 F_0 期望值分值. 由它们构成的线性秩统计量均可用于上述二样本位置问题的检验.

我们比较这些线性秩统计量，在各种典型总体下，求它们的渐近相对效率. 为了能应用本章第 1 节求渐近相对效率的基本定理 Noether 定理，需研究线性秩统计量在 H_1 下的渐近性质. 许多学者讨论过这个问题. Chernoff 和 Savage 首先于 1958 年在文章(Asymptotic normality and efficiency of certain nonparametric test statistics, *Ann. Math. Statist.* Vol 29, 972～994)中证明了 $c(i)$取值 0,1 时的线性秩统计量的 H_1 下的渐近正态性，其后 Govindarajulu 等和 Hájek 等作了许多改进. 这里叙述一个适合我们讨论的线性秩统计的结果，它是 Chernoff 和 Savage 的结果的一个特殊情况，其证明读者可参阅上述原始文献或陈希孺著《数理统计引论》.

设随机变量 $X_1, \cdots, X_m$ 相互独立同分布，分布函数为 $F(x)$；随机变量 $Y_1, \cdots, Y_n$ 相互独立同分布，分布函数为 $G(x)$，$F(x)$，$G(x)$连续，线性秩统计量

$$S_N = \sum_{i=1}^{n} a_N(R_i),$$

其中 $N = m + n$，$(Q_1, \cdots, Q_m, R_1, \cdots, R_n)$是混合样本 $X_1, \cdots, X_m$，$Y_1, \cdots, Y_n$ 对应的秩向量，分值

$$a_N(i) = b_N \phi\left(\frac{i}{N+1}\right) + d_N, \quad i = 1, \cdots, N,$$

其中 $\phi\left(\frac{i}{N+1}\right)$ 是函数 $\phi(u)$ 在 $\frac{i}{N+1}$ $(i=1,\cdots,N)$ 处的取值，$\phi(u)$ 是区间 $(0,1)$ 上的平方可积分值函数. 又记 $\frac{m}{N}=\lambda_N$，当 $N\to\infty$ 时，$\lambda_N\to\lambda$ $(0<\lambda<1)$，$F_m(x)$ 为 X 样本的经验分布函数，$G_n(x)$ 为 Y 样本的经验分布函数，则混合样本的经验分布函数为

$$H_N(x) = \lambda_N F_m(x) + (1-\lambda_N)G_n(x).$$

容易验证有

$$\begin{aligned} S_N &= \sum_{i=1}^{n} a_N(R_i) = b_N\sum_{i=1}^{n}\phi\left(\frac{R_i}{N+1}\right) + nd_N \\ &= nb_N\int_{-\infty}^{+\infty}\phi\left(\frac{N}{N+1}H_N(x)\right)\mathrm{d}G_n(x) + nd_N. \end{aligned}$$

Chernoff 和 Savage 的结果指明：统计量

$$T_N = \int_{-\infty}^{+\infty}\phi\left(\frac{N}{N+1}H_N(x)\right)\mathrm{d}G_n(x)$$

取正则化常数

$$\mu_N = \int_{-\infty}^{+\infty}\phi(H_N(x))\mathrm{d}G(x), \tag{5.3.4}$$

$$\begin{aligned} \sigma_N^2 = \frac{2}{N}\frac{\lambda_N}{1-\lambda_N}\bigg\{ &(1-\lambda_N)\iint_D \tilde{G}(x,y)\mathrm{d}F(x)\mathrm{d}F(y) \\ &+ \lambda_N\iint_D \tilde{F}(x,y)\mathrm{d}G(x)\mathrm{d}G(y)\bigg\}, \end{aligned} \tag{5.3.5}$$

其中

$$D = \{(x,y)\mid -\infty < x < y < \infty\},$$

$$\tilde{G}(x,y) = G(x)[1-G(y)]\phi'(H_N(x))\phi'(H_N(y)),$$

$$\tilde{F}(x,y) = F(x)[1-F(y)]\phi'(H_N(x))\phi'(H_N(y)).$$

$\phi(u)$ 在 $(0,1)$ 区间内连续，二阶导数存在，且存在常数 k 及 $\eta>0$，$\delta>0$，使

$$|\phi^{(i)}(u)| \leqslant K[u(1-u)]^{-i-\frac{1}{2}+\delta}, \quad i=0,1,2, 0<u<1,$$

$$N\sigma_N^2 \geqslant \eta, \quad 对一切 N,$$

这里 $\phi^{(0)}(u)=\phi(u)$，$\phi^{(1)}(u)$，$\phi^{(2)}(u)$ 为 $\phi(u)$ 的一阶、二阶导数，则

$$\frac{T_N-\mu_N}{\sigma_N} \text{ 有渐近正态分布 } N(0,1).$$

更特别当 $G(x)=F(x-\Delta)$，而 $N\to\infty$ 时，$\Delta=\Delta_N=\frac{c}{\sqrt{n}}$，其中 $c>0$ 为常数，对这一常见的特殊情况，在 Hájek 和 Šidák 的著作 (Theory of Rank Tests, Academic Press, 1967) 中指明：在 H_0 及 H_1 下，当

$$\int_0^1 \phi(u)\phi(u,f)\mathrm{d}u>0,$$

其中

$$\phi(u,f)=\frac{-f'(F^{-1}(u))}{f(F^{-1}(u))},$$

统计量 T_N 的渐近正态分布 $N(0,1)$ 的正则化常数有下列统一的形式

$$\mu_{T_N}(\Delta)=\bar{\phi}+\Delta\lambda_N\int_0^1 \phi(u)\phi(u,f)\mathrm{d}u, \tag{5.3.6}$$

$$\sigma_{T_N}(\Delta)=\frac{1}{N}\frac{\lambda_N}{1-\lambda_N}\int_0^1[\phi(u)-\bar{\phi}]^2\mathrm{d}u, \tag{5.3.7}$$

其中 $\bar{\phi}=\int_0^1 \phi(u)\mathrm{d}u$.

把 Chernoff 和 Savage 的正则化常数 μ_N，σ_N^2 与 Hájek 和 Šidák 的正则化常数 μ_{T_N}，$\sigma_{T_N}^2$ 进行比较，表达式的形式差异很大，但在给定的条件下，可用后者替代前者. 这个论证大致可按下述步骤进行. 首先注意到 $G(x)=F(x-\Delta)$，且 $\Delta=\Delta_N\to 0$ 时，有

$$\begin{aligned}
N\frac{1-\lambda_N}{\lambda_N}\sigma_N^2 &\approx 2\iint\limits_{0<x<y<1} x(1-y)\phi'(x)\phi'(y)\mathrm{d}x\mathrm{d}y \\
&= 2\iint\limits_{0<x<y<1} x\phi'(x)\phi'(y)\mathrm{d}x\mathrm{d}y \\
&\quad -\int_0^1\int_0^1 xy\phi'(x)\phi'(y)\mathrm{d}x\mathrm{d}y \\
&= 2\int_0^1\left[\int_x^1 \phi'(y)\mathrm{d}y\right]x\phi'(x)\mathrm{d}x-\left[\int_0^1 x\phi'(x)\mathrm{d}x\right]^2
\end{aligned}$$

$$= 2\phi(1)\int_0^1 x\mathrm{d}\phi(x) - \int_0^1 x\mathrm{d}\phi^2(x) - \left[\int_0^1 x\mathrm{d}\phi(x)\right]^2$$

$$= \int_0^1 \phi^2(x)\mathrm{d}x - \left[\int_0^1 \phi(x)\mathrm{d}x\right]^2$$

$$= \int_0^1 [\phi(x) - \overline{\phi}]^2\mathrm{d}x,$$

$$\mu_N = \int_{-\infty}^{+\infty} \phi(H_N(x))\mathrm{d}F(x-\Delta)$$

$$\approx \int_{-\infty}^{+\infty} [\phi(F(x-\Delta))$$

$$+ \lambda_N(F(x) - F(x-\Delta))\phi'(F(x-\Delta))]\mathrm{d}F(x-\Delta)$$

$$\approx \int_0^1 \phi(u)\mathrm{d}u + \lambda_N\int_0^1 [F(F^{-1}(u)-\Delta) - u]\phi'(u)\mathrm{d}u$$

$$\approx \overline{\phi} + \lambda_N\int_0^1 [-\Delta f(F^{-1}(u))]\phi'(u)\mathrm{d}u$$

$$= \overline{\phi} + \lambda_N\Delta\int_0^1 \phi(u)\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}\mathrm{d}u = \mu_{T_N}(\Delta).$$

进一步要证明当 $\Delta = \Delta_N = \dfrac{c}{\sqrt{N}}$ 的情形下，有

$$\sqrt{N}(\mu_N - \mu_{T_N}(\Delta)) \to 0, \quad 当\ N \to \infty.$$

这样根据 Slutsky 定理即知 T_N 的正则化常数可取 $\mu_{T_N}(\Delta)$，$\sigma^2_{T_N}(\Delta)$. 相应地，在 $G(x) = F(x-\Delta)$，且 $N \to \infty$ 时，$\Delta = \Delta_N = \dfrac{c}{\sqrt{N}}$ 的情形下，统计量 S_N 的渐近正态分布的正则化常数，在 H_0 及 H_1 下，有统一的形式：

$$\mu_{S_N}(\Delta) = n\,\overline{a}_N + \frac{\Delta mnb_N}{N}\int_0^1 \phi(u)\phi(u,f)\mathrm{d}u, \qquad (5.3.8)$$

$$\sigma_{S_N}(\Delta) = \frac{mnb_N^2}{N}\int_0^1 [\phi(u) - \overline{\phi}]^2\mathrm{d}u. \qquad (5.3.9)$$

上述 $\sigma_{S_N}(\Delta)$ 是不依赖 Δ 的常数，而 $\mu_{S_N}(\Delta)$ 是 Δ 的线性函数，容易验证本章第 1 节 Noether 定理的条件 $(A_1) \sim (A_6)$ 均能满足，故当

$$\int_0^1 \phi(u)\phi(u,f)\mathrm{d}u > 0$$

时，有效力因子

$$\mathrm{eff}(S_N) = \lim_{N\to\infty} \frac{\mu'_{S_N}(0)}{\sqrt{N\sigma_{S_N}(0)}} = \sqrt{\lambda(1-\lambda)}\,\frac{\int_0^1 \phi(u)\phi(u,f)\mathrm{d}u}{\sqrt{\int_0^1 [\phi(u)-\bar{\phi}]^2\mathrm{d}u}}. \tag{5.3.10}$$

例 1 由平方可积分值函数

$$\phi_L(u) = \begin{cases} u - 1/4, & u < 1/4, \\ 0, & 1/4 \leqslant u \leqslant 3/4, \\ u - 3/4, & u > 3/4 \end{cases}$$

产生线性秩统计量

$$S_L = \sum_{i=1}^{n} a_N(R_i),$$

其中

$$a_N(i) = (N+1)\phi_L\left(\frac{i}{N+1}\right)$$

$$= \begin{cases} i - \dfrac{N+1}{4}, & i < \dfrac{N+1}{4}, \\ 0, & \dfrac{N+1}{4} \leqslant i \leqslant \dfrac{3(N+1)}{4}, \\ i - \dfrac{3(N+1)}{4}, & i > \dfrac{3(N+1)}{4}, \end{cases}$$

则

$$\bar{\phi} = \int_0^1 \phi_L(u)\mathrm{d}u = 0,$$

$$\int_0^1 [\phi_L(u) - \bar{\phi}]^2\mathrm{d}u = \int_0^1 \phi_L^2(u)\mathrm{d}u = \frac{1}{96},$$

$$\int_0^1 \phi_L(u)\phi(u,f)\mathrm{d}u$$

$$=-\int_0^{1/4}\left(u-\frac{1}{4}\right)\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}\mathrm{d}u-\int_{3/4}^1\left(u-\frac{3}{4}\right)\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}\mathrm{d}u$$

$$=-\int_{\xi_0}^{\xi_{1/4}}\left(F(t)-\frac{1}{4}\right)f'(t)\mathrm{d}t-\int_{\xi_{3/4}}^{\xi_1}\left(F(t)-\frac{3}{4}\right)f'(t)\mathrm{d}t,$$

其中 ξ_p 为分布函数 $F(x)$的 p 分位数，即 $F(\xi_p)=p$. 若假定 $F(x)$ 的密度函数 $f(x)$在区间$[\xi_0,\xi_1]$外为 0，且

$$\lim_{x\downarrow\xi_0}f(x)=0,\quad \lim_{x\uparrow\xi_1}f(x)=0,$$

则上式经分部积分后，得

$$\int_0^1\phi_L(u)\phi(u,f)\mathrm{d}u=\int_{\xi_0}^{\xi_{1/4}}f^2(t)\mathrm{d}t+\int_{\xi_{3/4}}^{\xi_1}f^2(t)\mathrm{d}t.$$

由此有效力因子

$$\mathrm{eff}(S_L)=\sqrt{96\lambda(1-\lambda)}\left[\int_{\xi_0}^{\xi_{1/4}}f^2(t)\mathrm{d}t+\int_{\xi_{3/4}}^{\xi_1}f^2(t)\mathrm{d}t\right].$$

如果有$\int_0^1\phi(u,f)\mathrm{d}u=0$，(这是对分布函数 $F(x)$的要求，例如对正态分布 $\phi(u,f)=F^{-1}(u)$，则$\int_0^1F^{-1}(u)\mathrm{d}u=0$，即为要求 $F(x)$ 的期望为 0.)则效力因子

$$\mathrm{eff}(S)=\sqrt{\lambda(1-\lambda)}\frac{\int_0^1\phi(u)\phi(u,f)\mathrm{d}u}{\sqrt{\int_0^1[\phi(u)-\bar{\phi}]^2\mathrm{d}u}}$$

$$=\sqrt{\lambda(1-\lambda)\int_0^1\phi^2(u,f)\mathrm{d}u}\frac{\int_0^1\phi(u)\phi(u,f)\mathrm{d}u}{\sqrt{\int_0^1\phi^2(u,f)\mathrm{d}u\cdot\int_0^1[\phi(u)-\bar{\phi}]^2\mathrm{d}u}}$$

$$=\sqrt{\lambda(1-\lambda)\int_0^1\phi^2(u,f)\mathrm{d}u}\frac{\mathrm{cov}[\phi(U),\phi(U,f)]}{\sqrt{\mathrm{var}\phi(U)\cdot\mathrm{var}\phi(U,f)}},$$

其中随机变量 U 有(0,1)均匀分布. 从上式可以看出，当

$$\phi(u)=\phi(u,f)=\frac{-f'(F^{-1}(u))}{f(F^{-1}(u))}$$

时，对应的线性秩统计量的渐近效力最大. 渐近效力值达到

$$\sqrt{\lambda(1-\lambda)\int_0^1 \phi^2(u,f)\mathrm{d}u},$$

其中，值

$$\int_0^1 \phi^2(u,f)\mathrm{d}u = \int_0^1 \left[\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}\right]^2 \mathrm{d}u$$

$$= \int_{-\infty}^{+\infty} \left[\frac{f'(t)}{f(t)}\right]^2 f(t)\mathrm{d}t \xlongequal{\text{def}} I(F) \qquad (5.3.11)$$

称为分布 $F(x)$ 的 **Fisher 信息量**.

下面给出一些线性秩统计量的渐近相对效率. 这些统计量为

$$W = \sum_{i=1}^{n} R_i,$$

对应的 $a(i)=i=(N+1)\phi\left(\frac{i}{N+1}\right)$，$\phi(u)=u$.

$$NS = \sum_{i=1}^{n} a_{NS}(R_i),$$

其中 $a_{NS}(i)=\Phi^{-1}\left(\frac{i}{N+1}\right)$，$\Phi(u)$ 为正态 $N(0,1)$ 分布的分布函数，对应的 $\phi(u)=\Phi^{-1}(u)$.

$$L = \sum_{i=1}^{n} a_L(R_i),$$

其中 $a_L(i)=(N+1)\phi_L\left(\frac{i}{N+1}\right)$，

$$\phi_L(u) = \begin{cases} u-\frac{1}{4}, & \text{当 } u<\frac{1}{4} \text{ 时}, \\ 0, & \text{当 } \frac{1}{4}\leqslant u\leqslant\frac{3}{4} \text{ 时}, \\ u-\frac{3}{4}, & \text{当 } u>\frac{3}{4} \text{ 时}. \end{cases}$$

$$H = \sum_{i=1}^{n} a_H(R_i),$$

其中 $a_H(i)=(N+1)\phi_H\left(\frac{i}{N+1}\right)$，

$$\phi_H(u)=\begin{cases}-\dfrac{1}{4}, & \text{当 } u<\dfrac{1}{4} \text{ 时},\\ u-\dfrac{1}{2}, & \text{当 } \dfrac{1}{4}\leqslant u\leqslant\dfrac{3}{4} \text{ 时},\\ \dfrac{1}{4}, & \text{当 } u>\dfrac{3}{4} \text{ 时}.\end{cases}$$

$$M=\sum_{i=1}^{n}a_M(R_i),$$

其中 $a_M(i)=(N+1)\phi_M\left(\dfrac{i}{N+1}\right)$,

$$\phi_M(u)=\begin{cases}0, & \text{当 } u\leqslant\dfrac{1}{2} \text{ 时},\\ 1, & \text{当 } u>\dfrac{1}{2} \text{ 时}.\end{cases}$$

$$RS=\sum_{i=1}^{n}a_{RS}(R_i),$$

其中 $a_{RS}(i)=(N+1)\phi_{RS}\left(\dfrac{i}{N+1}\right)$,

$$\phi_{RS}(u)=\begin{cases}u-\dfrac{1}{2}, & \text{当 } u\leqslant\dfrac{1}{2} \text{ 时},\\ 0, & \text{当 } u>\dfrac{1}{2} \text{ 时}.\end{cases}$$

利用公式

$$\text{eff}(S)=\sqrt{\lambda(1-\lambda)}\,\frac{\int_0^1\phi(u)\phi(u,f)\mathrm{d}u}{\sqrt{\int_0^1[\phi(u)-\bar{\phi}]^2\mathrm{d}u}}$$

及

$$\text{ARE}(S_1,S_2)=\left[\frac{\text{eff}(S_1)}{\text{eff}(S_2)}\right]^2$$

可以算得下表.

分　布	ARE(NS,W)	ARE(L,W)	ARE(H,W)	ARE(M,W)	ARE(RS,W)
正　态	1.047	0.927	0.870	0.667	0.800
Logistic	0.955	0.781	0.945	0.750	0.800
双指数	0.847	0.500	1.125	1.333	0.800
Cauchy	0.708	0.264	1.339	1.333	0.800
指　数	~	2.000	0.500	0.333	1.800

从表中的数字可以看出：NS 是正态分位数分值统计量，对正态分布有较高的效力. 统计量 L 强调极端值，对轻尾分布有较高的效力. 统计量 H 抑低极端值，对重尾分布有较高的效力. 统计量 M 使用中位数分值，对重尾分布有较高的效力. 统计量 RS 强调值小的秩，对右斜分布有较高的效力. 而统计量 W 对 Logistic 分布最有效，对其他对称分布也有不错的效力. 对不同的总体分布，选用适宜的统计量可增加检验的功效. 有人建议采用下述适应性检验. 根据样本首先计算

$$Q_3=\frac{\overline{U}_{0.05}-\overline{M}_{0.50}}{\overline{M}_{0.50}-\overline{L}_{0.05}},\quad Q_4=\frac{\overline{U}_{0.05}-\overline{L}_{0.05}}{\overline{U}_{0.50}-\overline{L}_{0.50}},$$

其中 $\overline{U}_p$ 为将样本按值的大小顺序排列，从高端取全样本量的比例为 p 的一部分的平均值. $\overline{M}_p$ 为从中段取全样本量的比例为 p 的一部分的平均值. $\overline{L}_p$ 为从低端取全样本量的比例为 p 的一部分的平均值. 从 Q_3 看总体分布倾斜的程度，从 Q_4 看分布轻重尾的情况，根据它们的值选用检验统计量：

当 $Q_4>7$ 时，用 M 统计量；

当 $Q_3>2, Q_4\leqslant 7$ 时，用 RS 统计量；

当 $Q_3\leqslant 2, 2\leqslant Q_4\leqslant 7$ 时，用 W 统计量；

当 $Q_3\leqslant 2, 1\leqslant Q_4<2$ 时，用 L 统计量.

§5.4　二样本尺度问题线性秩统计量的渐近相对效率

二样本尺度问题：分布 $F(t)$ 连续，样本 $X_1,\cdots,X_m$ 相互独立

同分布，分布函数为 $F(x-\theta)$；$Y_1,\cdots,Y_n$ 相互独立同分布，分布函数为 $F\left(\frac{x-\theta}{\eta}\right)$，$N=m+n$. 检验假设

$$H_0:\ \eta=1,\qquad H_1:\ \eta>1.$$

如果记 $\eta=\mathrm{e}^{\Delta}$，则检验假设

$$H_0:\ \Delta=0,\qquad H_1:\ \Delta>0$$

可用线性秩统计量

$$S_N=\sum_{i=1}^{n}a_N(R_i),$$

其中 $R=(Q_1,\cdots,Q_m,R_1,\cdots,R_n)$ 仍是混合样本 $(X_1,\cdots,X_m,Y_1,\cdots,Y_n)$ 的秩向量. 在 H_0 下，R 在

$$\mathscr{R}=\{(1,\cdots,N)\ \text{的全部排列}\}$$

上均匀分布. 从直观上看只要当 $i\leqslant\frac{N+1}{2}$ 时，$a_N(i)$ 非升；当 $i\geqslant\frac{N+1}{2}$ 时，$a_N(i)$ 非降，那么对应的统计量 S_N，当取大的值时有利于 H_1，应否定 H_0. 例如可取下列线性秩统计量

$$S_M=\sum_{i=1}^{n}a_M(R_i),$$

其中 $a_M(i)=(N+1)^2\phi\left(\frac{i}{N+1}\right)$，$\phi(u)=\left(u-\frac{1}{2}\right)^2$.

$$S_{AB}=\sum_{i=1}^{n}a_{AB}(R_i),$$

其中 $a_{AB}(i)=(N+1)\phi\left(\frac{i}{N+1}\right)$，$\phi(u)=\left|u-\frac{1}{2}\right|$.

$$S_K=\sum_{i=1}^{n}a_K(R_i),$$

其中 $a_K(i)=\left[\phi\left(\frac{i}{N+1}\right)\right]^2$，$\phi(u)=\Phi^{-1}(u)$，$\Phi(u)$ 为标准正态分布函数.

与位置问题类似，Hájek 等在其著作中指明，当 $N\to\infty$，$\frac{m}{N}\to$

$\lambda>0$，而取 $\Delta=\Delta_N=\frac{c}{\sqrt{N}}$，$c>0$ 为常数时，在 H_0,H_1 下，此类线性秩统计量渐近正态分布 $N(0,1)$ 的正则化常数，有统一的形式

$$\mu_{S_N}(\Delta)=n\,\bar{a}_N+\frac{\Delta mnb_N}{N}\int_0^1\phi(u)\phi_{SC}(u,f)\mathrm{d}u, \quad (5.4.1)$$

$$\sigma_{S_N}^2(\Delta)=\frac{mnb_N^2}{N}\int_0^1[\phi(u)-\bar{\phi}]^2\mathrm{d}u, \quad (5.4.2)$$

其中，$a_N(i)=b_N\phi\left(\frac{i}{N+1}\right)+d_N$，$\phi(u)$ 为平方可积分值函数，

$$\phi_{SC}(u,f)=-1-F^{-1}(u)\frac{f'[F^{-1}(u)]}{f[F^{-1}(u)]}.$$

当

$$\int_0^1\phi(u)\phi_{SC}(u,f)\mathrm{d}u>0$$

时，有

$$\mathrm{eff}(S_N)=\sqrt{\lambda(1-\lambda)}\frac{\int_0^1\phi(u)\phi_{SC}(u,f)\mathrm{d}u}{\sqrt{\int_0^1[\phi(u)-\bar{\phi}]^2\mathrm{d}u}}. \quad (5.4.3)$$

与位置问题时情况类似，当 $\int_0^1\phi_{SC}(u,f)\mathrm{d}u=0$ 时，取 $\phi(u)=\phi_{SC}(u,f)$，则 $\mathrm{eff}(S_N)$ 达最大值. $\phi(u)=\phi_{SC}(u,f)$ 是针对总体分布 $F(x)$ 的二样本尺度问题的最优分值函数.

用上述公式可计算得下表

分　布	ARE(*AB*,*M*)	ARE(*K*,*M*)
正　态	0.800	1.316
Logistic	0.837	1.116
双指数	0.864	1.052
Cauchy	1.067	0.598
指　数	0.806	~

§5.5 一样本位置问题线性符号秩统计量的渐近相对效率

一样本位置问题：样本 $X_1,\cdots,X_n$ 相互独立同分布，分布为 $F(x-\theta)$，分布函数 $F(t)$ 关于 0 点对称，有密度函数 $f(t)$，检验假设

$$H_0:\theta=0,\qquad H_1:\theta>0.$$

检验此问题可以用线性符号秩统计量(参见第三章第 5 节)

$$S_n^+=\sum_{i=1}^{n}\Psi_i a_n(R_i^+),$$

其中 $R^+=(R_1^+,\cdots,R_n^+)$ 是样本 $(X_1,\cdots,X_n)$ 对应的绝对秩向量，$\{a_n(i)\}$ 为分值，Ψ_i 为 X_i 的符号统计量，即

$$\Psi_i=\begin{cases}0, & \text{当 } X_i\leqslant 0 \text{ 时},\\ 1, & \text{当 } X_i>1 \text{ 时},\end{cases}\quad i=1,\cdots,n.$$

我们可以看到当 $0\leqslant a_n(1)\leqslant\cdots\leqslant a_n(n)$，$a_n(n)>0$ 时，则统计量 S_n^+ 取大的值有利于 H_1，否定 H_0. 例如有下列线性符号秩统计量均可以用于检验一样本位置问题. 我们仍假定

$$a_n(i)=b_n\phi^+\left(\frac{i}{n+1}\right),$$

这里 $\phi^+(\cdot)$ 是非负非降平方可积分值函数，不依赖于 n.

$$W^+=\sum_{i=1}^{n}\Psi_i R_i^+,$$

对应的 $a_n(i)=i$，$\phi^+(u)=u$，$a_n(i)=(n+1)\phi^+\left(\frac{i}{n+1}\right)$.

$$B=\sum_{i=1}^{n}\Psi_i,$$

这里 $a_n(i)\equiv1$，$\phi^+(u)\equiv1$.

$$H^+=\sum_{i=1}^{n}\Psi_i a_H(R_i^+),$$

这里
$$\phi_H^+(u)=\begin{cases} u, & \text{当 } 0<u\leqslant\frac{1}{2}\text{ 时},\\ \frac{1}{2}, & \text{当 }\frac{1}{2}<u<1\text{ 时},\end{cases}$$

$$a_H(i)=(n+1)\phi_H^+\left(\frac{i}{n+1}\right)=\begin{cases} i, & \text{当 } i\leqslant\frac{n+1}{2}\text{ 时},\\ \frac{n+1}{2}, & \text{当 } i>\frac{n+1}{2}\text{ 时}.\end{cases}$$

$$L^+=\sum_{i=1}^{n}\Psi_i a_L(R_i^+),$$

这里

$$\phi_L^+(u)=\begin{cases} 0, & \text{当 } u\leqslant\frac{1}{2}\text{ 时},\\ u-\frac{1}{2}, & \text{当 } u>\frac{1}{2}\text{ 时},\end{cases}$$

$$a_L(i)=(n+1)\phi_L^+\left(\frac{i}{n+1}\right)=\begin{cases} 0, & \text{当 } i\leqslant\frac{n+1}{2}\text{ 时},\\ i-\frac{n+1}{2}, & \text{当 } i>\frac{n+1}{2}\text{ 时}.\end{cases}$$

$$NS^+=\sum_{i=1}^{n}\Psi_i a_{NS}(R_i^+),$$

这里

$$\phi_{NS}^+(u)=\Phi^{-1}\left(\frac{1}{2}+\frac{1}{2}u\right),$$

$\Phi^{-1}(u)$是标准正态分布函数的反函数,

$$a_{NS}(i)=\phi_{NS}^+\left(\frac{i}{n+1}\right)=\Phi^{-1}\left(\frac{1}{2}+\frac{1}{2}\,\frac{i}{n+1}\right).$$

在 $n\to\infty$,求它们的渐近相对效率. 与二样本位置问题类似,在假设 H_0: $\theta=0$ 和 H_1: $\theta=\frac{c}{\sqrt{n}}$ ($c>0$ 固定常数)下,统计量 S_n^+ 渐近正态 $N(0,1)$分布的正则化常数可以有形式:

$$\mu_{S_n^+}(\theta)=\frac{n}{2}\,\overline{a}_n+\frac{\theta n b_n}{2}\int_0^1\phi^+(u)\phi^+(u,f)\mathrm{d}u, \quad (5.5.1)$$

$$\sigma^2_{S_n^+}(\theta) = \frac{nb_n^2}{4}\int_0^1[\phi^+(u)]^2\mathrm{d}u, \qquad (5.5.2)$$

其中
$$\bar{a}_n = \frac{1}{n}\sum_{i=1}^{n} a_n(i),$$

$$\phi^+(u,f) = \frac{-f'\left[F^{-1}\left(\frac{1}{2}+\frac{1}{2}u\right)\right]}{f\left[F^{-1}\left(\frac{1}{2}+\frac{1}{2}u\right)\right]}.$$

它们是 θ 的线性函数,由本章第 1 节 Noether 定理可得效力因子

$$\mathrm{eff}(S_n^+) = \frac{\int_0^1\phi^+(u)\phi^+(u,f)\mathrm{d}u}{\sqrt{\int_0^1[\phi^+(u)]^2\mathrm{d}u}}. \qquad (5.5.3)$$

根据上式,可以算得

$$\mathrm{eff}(W^+) = \sqrt{12}\int_{-\infty}^{+\infty}f^2(x)\mathrm{d}x;$$

$$\mathrm{eff}(B) = 2f(0);$$

$$\mathrm{eff}(H^+) = 4\sqrt{6}\int_0^{\xi_{3/4}}f^2(x)\mathrm{d}x,$$

其中 ξ_p 为总体分布 $F(x)$的 p 分位数,

$$\mathrm{eff}(L^+) = \sqrt{96}\left[\int_{\xi_0}^{\xi_{1/4}}f^2(x)\mathrm{d}x + \int_{\xi_{3/4}}^{\xi_1}f^2(x)\mathrm{d}x\right],$$

$$\mathrm{eff}(NS^+) = -\int_0^1\Phi^{-1}(u)\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}\mathrm{d}u.$$

例 1 作为例子计算 $\mathrm{eff}(H^+)$.

设 $F(x)$有密度 $f(x)$,关于 0 点对称.

$$\phi_H^+(u) = \begin{cases} u, & \text{当 } u \leqslant \frac{1}{2} \text{ 时}, \\ \frac{1}{2}, & \text{当 } u > \frac{1}{2} \text{ 时}, \end{cases}$$

则

$$\int_0^1 [\phi_H^+(u)]^2 \mathrm{d}u = \frac{1}{6},$$

$$\int_0^1 [\phi_H^+(u)]\phi^+(u,f)\mathrm{d}u$$

$$= \int_0^{1/2} u \left[\frac{-f'\left[F^{-1}\left(\frac{1}{2}+\frac{1}{2}u\right)\right]}{f\left[F^{-1}\left(\frac{1}{2}+\frac{1}{2}u\right)\right]} \right] \mathrm{d}u$$

$$+ \int_{1/2}^1 \frac{1}{2} \left[\frac{-f'\left[F^{-1}\left(\frac{1}{2}+\frac{1}{2}u\right)\right]}{f\left[F^{-1}\left(\frac{1}{2}+\frac{1}{2}u\right)\right]} \right] \mathrm{d}u$$

$$= -2\int_{\xi_{1/2}}^{\xi_{3/4}} [2F(t)-1]f'(t)\mathrm{d}t - \int_{\xi_{3/4}}^{\xi_1} f'(t)\mathrm{d}t$$

$$= -2[f(t)[2F(t)-1]]_{\xi_{1/2}}^{\xi_{3/4}} + 4\int_{\xi_{1/2}}^{\xi_{3/4}} f^2(t)\mathrm{d}t - [f(t)]_{\xi_{3/4}}^{\xi_1}$$

$$= -f(\xi_{3/4}) + 4\int_{\xi_{1/2}}^{\xi_{3/4}} f^2(t)\mathrm{d}t - f(\xi_1) + f(\xi_{3/4}).$$

注意到应有 $f(\xi_1)=0$,所以

$$\int_0^1 \phi_H^+(u)\phi^+(u,f)\mathrm{d}u = 4\int_{\xi_{1/2}}^{\xi_{3/4}} f^2(t)\mathrm{d}t = 4\int_0^{\xi_{3/4}} f^2(t)\mathrm{d}t.$$

所以

$$\mathrm{eff}(H^+) = 4\sqrt{6}\int_0^{\xi_{3/4}} f^2(t)\mathrm{d}t.$$

上面给出的一些线性符号秩统计量之间的渐近相对效率有下表.

分 布	ARE(B,W^+)	ARE(H^+,W^+)	ARE(L^+,W^+)	ARE(NS^+,W^+)
正 态	0.667	0.870	0.927	1.047
Logistic	0.750	0.945	0.781	0.995
双指数	1.333	1.125	0.500	0.847
Cauchy	1.333	1.339	0.264	0.708

习　题

1. 对假设检验问题 H_0 和备择假设 H_1 有三个检验统计量 S，T 和 V，说明

$$\mathrm{ARE}(S,V)=\mathrm{ARE}(S,T)\cdot\mathrm{ARE}(T,V)=\frac{1}{\mathrm{ARE}(V,S)}.$$

2. 设 $\{S_n\}$ 和 $\{T_n\}$ 是关于零假设 H_0 和备择假设 H_1 的两列检验统计量，假设存在正则化系数 $\{\mu_n(\theta)\}$ 和 $\{\sigma_n(\theta)\}$ 满足 Noether 定理的条件 $(A_3)\sim(A_6)$，且 $\dfrac{S_n-\mu_n(\theta)}{\sigma_n(\theta)}$ 满足定理的条件 (A_1)，(A_2). 又

$$\frac{S_n-T_n}{\sigma_n(\theta)}\xrightarrow{P}0,\quad \text{对 } \theta \text{ 一致收敛}.$$

说明基于 S_n 和基于 T_n 的检验有相同的效力.

3. 证明第 2 节定理 2.4.

4. 验证第 2 节关于 $\mathrm{ARE}(W^+,T)$，$\mathrm{ARE}(B,T)$ 的表.

5. 具体计算二样本问题的 $\mathrm{eff}(W)$ 和 $\mathrm{eff}(T)$.

6. 利用定理 2.4 验证 Wilcoxon 秩和统计量在 H_1 下有渐近正态分布.

7. 分布函数 $F_0(\cdot)$ 分别是下列分布：

(a) $(0,1)$ 上均匀分布；

(b) $(0,\infty)$ 上指数分布，$f(x)=\mathrm{e}^{-x}$，当 $x>0$.

找出对应的 F_0 期望值分值 $a^*_{F_0}(i)$ 和 F_0 分位数分值 $a_{F_0}(i)$.

8. 验证第 3 节中关于 $\mathrm{ARE}(L,W)$，$\mathrm{ARE}(NS,W)$，$\mathrm{ARE}(H,W)$，$\mathrm{ARE}(M,W)$，$\mathrm{ARE}(RS,W)$ 的表.

9. 对第 4 节中的统计量 S_M，说明当总体分布 $F(x)$ 的中位数为 0 时，

$$\mathrm{eff}(S_M)=\sqrt{720\lambda(1-\lambda)}\int_{-\infty}^{+\infty}\left[F(x)-\frac{1}{2}\right]xf(x)\mathrm{d}x.$$

10. 验证第 5 节关于 $\mathrm{ARE}(L^+,W^+)$，$\mathrm{ARE}(NS^+,W^+)$ 的表.

第六章　由经验分布产生的非参数估计

总体分布的分位数和分位数的函数的估计是非参数估计的基本内容.这类估计一般不假定总体分布的具体形式.估计涉及的基本统计量是样本的次序统计量和经验分布.

§6.1　次序统计量的分布

设随机变量 $X_1,\cdots,X_n$ 相互独立同分布,是来自一个连续分布总体 $F(x)$ 的样本,总体密度函数记为 $f(x)$.令 $X_{(i)}$ 是这一样本观测从小到大排列的第 i 个最小者,有 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$. $X_{(1)},\cdots,X_{(n)}$ 称为样本 $X_1,\cdots,X_n$ 的**次序统计量**.与样本不同,它们相互间既不独立也不同分布.

定理 1.1　设 $X_{(1)},\cdots,X_{(n)}$ 是连续分布总体 $F(x)$(密度函数为 $f(x)$)的样本次序统计量,则 $(X_{(1)},\cdots,X_{(n)})$ 的联合密度函数为

$$g(x_{(1)},\cdots,x_{(n)})=\begin{cases}n!\prod\limits_{i=1}^{n}f(x_{(i)}), & \text{当 } -\infty<x_{(1)}<\cdots<x_{(n)}<\infty \text{ 时},\\ 0, & \text{其他}.\end{cases}\tag{6.1.1}$$

证明　由于总体分布连续,不妨认为样本 $X_1,\cdots,X_n$ 的取值空间为

$$\{(x_1,\cdots,x_n)\mid -\infty<x_i<\infty,x_i\neq x_j,i\neq j,i,j=1,\cdots,n\}.$$

将这一空间分割成 $n!$ 块,每一块对应于 $(1,\cdots,n)$ 的一个排列 $(i_1,\cdots,i_n)$.与 $(i_1,\cdots,i_n)$ 对应的块为

$$\{(x_1,\cdots,x_n)\mid -\infty<x_{i_1}<\cdots<x_{i_n}<\infty\}.$$

在此块上令 $x_{i_1}=x_{(1)},\cdots,x_{i_n}=x_{(n)}$，则变换的 Jacobian 行列式为 1，密度 $f(x_1)\cdots f(x_n)$ 变换后经整理次序，即为 $f(x_{(1)})\cdots f(x_{(n)})$. 故当 $x_{(1)}<\cdots<x_{(n)}$ 时，$(X_{(1)},\cdots,X_{(n)})$ 的联合密度函数为

$$g(x_{(1)},\cdots,x_{(n)})=\sum_{(i_1,\cdots,i_n)}f(x_{(1)})\cdots f(x_{(n)})=n!\prod_{i=1}^{n}f(x_{(i)}),$$

其他情形为 0 是显然的. 证毕.

例 1 若总体分布为(0,1)均匀分布，则样本次序统计量 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 的联合密度函数为

$$g(x_{(1)},\cdots,x_{(n)})=\begin{cases}n!, & \text{当} -\infty<x_{(1)}<\cdots<x_{(n)}<\infty \text{ 时},\\ 0, & \text{其他}.\end{cases}$$

若总体分布为指数分布，密度 $f(x)=\mathrm{e}^{-x}$，$0<x<\infty$，则样本次序统计量的联合密度函数为

$$g(x_{(1)},\cdots,x_{(n)})$$
$$=\begin{cases}n!\ \exp\left\{-\sum_{i=1}^{n}x_{(i)}\right\}, & \text{当 } 0<x_{(1)}<\cdots<x_{(n)}<\infty \text{ 时},\\ 0, & \text{其他}.\end{cases}$$

定理 1.2 设 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 是连续分布总体 $F(x)$（密度函数 $f(x)$）的样本次序统计量，则当 $x_{(i)}<x_{(j)}$ 时，$(X_{(i)},X_{(j)})$ $(1\leqslant i<j\leqslant n)$ 的二维边缘密度函数为

$$\begin{aligned}g_{ij}(x_{(i)},x_{(j)})=&\frac{n!}{(i-1)!(j-i-1)!(n-j)!}\\&\cdot\{F(X_{(i)})\}^{i-1}\{F(x_{(j)})-F(x_{(i)})\}^{j-i-1}\\&\cdot\{1-F(x_{(j)})\}^{n-j}f(x_{(i)})f(x_{(j)}).\qquad(6.1.2)\end{aligned}$$

证明 由定理 1.1 之联合密度函数对变量 $(x_{(1)},\cdots,x_{(i-1)})$，$(x_{(i+1)},\cdots,x_{(j-1)})$，$(x_{(j+1)},\cdots,x_{(n)})$ 积分即得 $(X_{(i)},X_{(j)})$ 的联合密度函数. 当 $x_{(i)}<x_{(j)}$ 时，

$$g_{ij}(x_{(i)},x_{(j)})=n!\ f(x_{(i)})f(x_{(j)})$$

$$\times\int_{-\infty}^{x_{(i)}}\cdots\int_{-\infty}^{x_{(2)}}f(x_{(1)})\cdots f(x_{(i-1)})\mathrm{d}x_{(1)}\cdots\mathrm{d}x_{(i-1)}$$

$$\times\int_{x_{(i)}}^{x_{(j)}}\cdots\int_{x_{(i)}}^{x_{(i+2)}}f(x_{(i+1)})\cdots f(x_{(j-1)})\mathrm{d}x_{(i+1)}\cdots\mathrm{d}x_{(j-1)}$$

$$\times\int_{x_{(j)}}^{\infty}\cdots\int_{x_{(j)}}^{x_{(j+2)}}f(x_{(j+1)})\cdots f(x_{(n)})\mathrm{d}x_{(j+1)}\cdots\mathrm{d}x_{(n)}.$$

而积分

$$\int_{-\infty}^{x_{(i)}}\cdots\int_{-\infty}^{x_{(2)}}f(x_{(1)})\cdots f(x_{(i-1)})\mathrm{d}x_{(1)}\cdots\mathrm{d}x_{(i-1)}=\frac{\{F(x_{(i)})\}^{i-1}}{(i-1)!},$$

$$\int_{x_{(i)}}^{x_{(j)}}\cdots\int_{x_{(i)}}^{x_{(i+2)}}f(x_{(i-1)})\cdots f(x_{(j-1)})\mathrm{d}x_{(i+1)}\cdots\mathrm{d}x_{(j-1)}$$

$$=\frac{\{F(x_{(j)})-F(x_{(i)})\}^{j-i-1}}{(j-i-1)!},$$

$$\int_{x_{(j)}}^{\infty}\cdots\int_{x_{(j)}}^{x_{(j+2)}}f(x_{(j+1)})\cdots f(x_{(n)})\mathrm{d}x_{(j+1)}\cdots\mathrm{d}x_{(n)}$$

$$=\frac{\{1-F(x_{(j)})\}^{n-j}}{(n-j)!}.$$

整理即得定理结论. 证毕.

定理 1.3 设 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 是连续分布总体 $F(x)$(密度函数 $f(x)$)的样本次序统计量,则 $X_{(i)}(1\leqslant i\leqslant n)$ 的一维边缘密度函数为

$$g_i(x_{(i)})=\frac{n!}{(i-1)!(n-i)!}\{F(x_{(i)})\}^{i-1}\{1-F(x_{(i)})\}^{n-i}f(x_{(i)}),$$
$$-\infty<x_{(i)}<\infty. \tag{6.1.3}$$

证明 与定理 1.2 类似.

例 2 当 $X_1,\cdots,X_n$ 独立同分布,来自连续分布总体 $F(x)$ 时,$F(X_1),\cdots,F(X_n)$ 可看作是来自 $(0,1)$ 均匀分布的样本,而 $F(X_{(1)}),\cdots,F(X_{(n)})$ 是上述均匀分布样本的次序统计量. 故 $(F(X_{(i)}),F(X_{(j)}))$ $(1\leqslant i<j\leqslant n)$ 的二维联合分布是二变量 Dirichlet 分布,密度函数为

$$g_{ij}(u,v)=\frac{n!}{(i-1)!\ (j-i-1)!\ (n-j)!}u^{i-1}(v-u)^{j-i-1}(1-v)^{n-j}$$
$$=\frac{\Gamma(n+1)}{\Gamma(i)\Gamma(j-i)\Gamma(n-j+1)}u^{i-1}(v-u)^{j-i-1}(1-v)^{n-j},$$
当 $0<u<v<1$ 时.

§6.2 分位数的估计

对一个总体分布 $F(x)$,其 p 分位数 ξ_p 是满足下述条件的一个数: $F(\xi_p)\geqslant p, F(\xi_p-0)\leqslant p\ (0<p<1)$,估计 ξ_p 的问题是常见的估计问题. 对参数分布族而言,p 分位数常可通过参数表出,因而估计参数后,可获得相应的 p 分位数的估计. 但在对总体分布形式无先验知识的情况下,用样本次序统计量可构成 p 分位数的非参数估计,即用样本的 p 分位数作为总体分布的 p 分位数的估计.

定义 2.1 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布,是来自总体 $F(x)$的样本,由 $X_1,\cdots,X_n$ 构成的经验分布函数记为 $F_n(x)$,则

$$F_n^{-1}(p)=\inf\{x: F_n(x)\geqslant p\} \tag{6.2.1}$$

称为**样本的 p 分位数**.

用样本的 p 分位数估计总体分布的 p 分位数有下列强相合性和渐近正态性定理.

定理 2.1 对 $0<p<1$,ξ_p 是满足 $F(x)\geqslant p, 1-F(x-0)\geqslant 1-p$的总体 $F(x)$的 p 分位数. 如果 ξ_p 是唯一的,则估计量

$$\hat{\xi}_p=F_n^{-1}(p)$$

是 ξ_p 的强相合估计.

证明 令 $\varepsilon>0$,由于 ξ_p 是总体 $F(x)$的唯一的 p 分位数,有

$$F(\xi_p-\varepsilon)<p<F(\xi_p+\varepsilon),$$

而由经验分布函数的性质,知对给定的 $\xi_p-\varepsilon$ 和 $\xi_p+\varepsilon$,$F_n(\xi_p-\varepsilon)$ 和 $F_n(\xi_p+\varepsilon)$分别概率为 1 地收敛到 $F(\xi_p-\varepsilon)$和 $F(\xi_p+\varepsilon)$,故当

$n\to\infty$时，

$$P\{F_m(\xi_p-\varepsilon)<p<F_m(\xi_p+\varepsilon),\text{对一切 } m>n\}\to 1.$$

从而由 $F_m(x)$的单调性，得当 $n\to\infty$时，

$$P\{\xi_p-\varepsilon<\hat{\xi}_p(m)\leqslant\xi_p+\varepsilon,\text{对一切 } m>n\}\to 1,$$

式中 $\hat{\xi}_p(m)$表示样本量为 m 时的样本 p 分位数. 这也就是

$$P\left\{\sup_{m\geqslant n}|\hat{\xi}_p(m)-\xi_p|\leqslant\varepsilon\right\}\to 1,\quad n\to\infty.$$

证毕.

定理 2.2 对 $0<p<1$，如果总体分布 $F(x)$的密度函数 $f(x)$在 ξ_p 的一个邻域内大于 0，且在 ξ_p 点连续，则样本分位数 $\hat{\xi}_p(n)$有渐近正态分布 $N(\xi_p,p(1-p)/nf^2(\xi_p))$.

证明 对任给的 t，记

$$G_n(t)=P\left\{\frac{\sqrt{n}\,(\hat{\xi}_p(n)-\xi_p)}{D}\leqslant t\right\},\tag{6.2.2}$$

其中 $D=\sqrt{p(1-p)}/f(\xi_p)$，有

$$\begin{aligned}G_n(t)&=P\{\hat{\xi}_p(n)\leqslant\xi_p+tn^{-\frac{1}{2}}D\}\\&=P\{p\leqslant F_n(\xi_p+tn^{-\frac{1}{2}}D)\}.\end{aligned}\tag{6.2.3}$$

而由经验分布函数的性质，知 $nF_n(\xi_p+tn^{-\frac{1}{2}}D)$服从参数为 $(n,F(\xi_p+tn^{-\frac{1}{2}}D))$的二项分布. 记 $\Delta_{nt}=F(\xi_p+tn^{-\frac{1}{2}}D)$，$Z_n(\Delta)$为参数$(n,\Delta)$的二项分布随机变量，由 Berry-Esseen 中心极限定理，存在不依赖 n 的常数 A，使

$$\begin{aligned}&\sup_{-\infty<x<\infty}\left|P\left\{\frac{Z_n(\Delta)-n\Delta}{[n\Delta(1-\Delta)]^{1/2}}\leqslant x\right\}-\Phi(x)\right|\\&\qquad\leqslant An^{-\frac{1}{2}}E|W|^3=An^{-\frac{1}{2}}\frac{(1-\Delta)^2+\Delta^2}{\sqrt{\Delta(1-\Delta)}},\quad n\to\infty,\end{aligned}$$

其中 $W=\dfrac{T-\Delta}{[\Delta(1-\Delta)]^{\frac{1}{2}}}$，$T$ 为参数 Δ 的(0,1)两点分布随机变量. 同时，有

$$G_n(t) = P\left\{ \frac{Z_n(\Delta_{nt}) - n\Delta_{nt}}{[n\Delta_{nt}(1-\Delta_{nt})]^{1/2}} \geqslant - \frac{n^{1/2}(\Delta_{nt} - p)}{[\Delta_{nt}(1-\Delta_{nt})]^{1/2}} \right\}. \tag{6.2.4}$$

若记 $C_{nt}=n^{1/2}(\Delta_{nt}-p)/[\Delta_{nt}(1-\Delta_{nt})]^{1/2}$，则

$$\begin{aligned} &|\Phi(t)-G_n(t)| \\ &\quad = |\Phi(t)-\Phi(-C_{nt})+(1-G_n(t))-(1-\Phi(-C_{nt}))| \\ &\quad \leqslant |(1-G_n(t))-\Phi(-C_{nt})|+|\Phi(t)-(1-\Phi(-C_{nt}))|. \end{aligned}$$

由于 $F(x)$ 在 ξ_p 连续，故 $n\to\infty$ 时，有

$$\Delta_{nt}(1-\Delta_{nt}) \to p(1-p), \quad n^{-\frac{1}{2}} \frac{(1-\Delta_{nt})^2+\Delta_{nt}^2}{[\Delta_{nt}(1-\Delta_{nt})]^{1/2}} \to 0.$$

由前述 Berry-Esseen 中心极限定理的不等式，知

$$|(1-G_n(t)) - \Phi(-C_{nt})| \to 0, \quad n\to\infty.$$

又因 $f(x)$ 在 ξ_p 连续，

$$\begin{aligned} C_{nt} &= \frac{n^{1/2}(\Delta_{nt}-p)}{[\Delta_{nt}(1-\Delta_{nt})]^{1/2}} = \frac{n^{1/2}[F(\xi_p+tn^{-1/2}D)] - F(\xi_p)}{[\Delta_{nt}(1-\Delta_{nt})]^{1/2}} \\ &\to \frac{tD}{[p(1-p)]^{1/2}} f(\xi_p) = t, \quad n\to\infty, \end{aligned}$$

因此，当 $n\to\infty$ 时，有

$$|\Phi(t) - (1-\Phi(-C_{nt}))| = |\Phi(t)-\Phi(C_{nt})| \to 0,$$

从而有定理之结论. 证毕.

例 1 当总体分布为 $[0,1]$ 上均匀分布，

$$F(x)=x, \quad 0\leqslant x\leqslant 1, \quad f(x)=1, \quad 0\leqslant x\leqslant 1,$$

$\xi_{1/2}=\frac{1}{2}$，密度函数在此点连续. 故样本中位数 $\xi_{1/2}$ 有渐近正态分布 $N\left(\frac{1}{2},\frac{1}{4n}\right)$. 而当总体分布为

$$F(x)=\begin{cases} x, & \text{当 } 0\leqslant x\leqslant \frac{1}{2} \text{ 时}, \\ 2x-\frac{1}{2}, & \text{当 } \frac{1}{2}< x\leqslant \frac{3}{4} \text{ 时}, \end{cases}$$

密度函数

$$f(x)=\begin{cases}1, & \text{当 } 0\leqslant x\leqslant\frac{1}{2}\text{ 时},\\ 2, & \text{当 }\frac{1}{2}<x\leqslant\frac{3}{4}\text{ 时},\end{cases}$$

$\xi_{1/2}=\frac{1}{2}$. 但密度函数在此点不连续，不满足定理条件. 然而从定理的证明可以看出对于 $t<0$ 和 $t>0$，C_n 分别趋于

$$[p(1-p)]^{-1/2}tDf(\xi_p-0) \text{ 和 } [p(1-p)]^{-1/2}tDf(\xi_p+0),$$

故仍可用正态近似计算出概率

$$P\left\{n^{1/2}\left(\hat{\xi}_{1/2}-\frac{1}{2}\right)\leqslant t\right\}.$$

当 $t<0$ 时，使用 $N\left(0,\frac{1}{4}\right)$，而当 $t>0$ 时，使用 $N\left(0,\frac{1}{16}\right)$.

§6.3 分位数的区间估计

设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，是来自连续分布 $F(x)$ 的样本，$X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 为样本次序统计量，则

$$U=F(X_{(k)}),\quad V=F(X_{(l)}),\quad k<l$$

的二维联合分布为 Dirichlet 分布，密度函数为

$$f(u,v)=\frac{\Gamma(n+1)}{\Gamma(k)\Gamma(l-k)\Gamma(n-l+1)}u^{k-1}(v-u)^{l-k-1}(1-v)^{n-l},$$

$$\text{当 } 0<u<v<1 \text{ 时}.$$

现在考虑不等式

$$F(X_{(k)})<p<F(X_{(l)}),$$

由于 $F(x)$ 连续单调，上述不等式等价于

$$X_{(k)}<\xi_p<X_{(l)},$$

因而有 $\quad P\{X_{(k)}<\xi_p<X_{(l)}\}=\int_p^1\int_0^p f(u,v)\mathrm{d}u\mathrm{d}v$

$$=\int_0^p\int_u^1 f(u,v)\mathrm{d}v\mathrm{d}u-\int_0^p\int_u^p f(u,v)\mathrm{d}v\mathrm{d}u$$

$$= \int_0^p \int_v^1 f(u,v)\mathrm{d}v\mathrm{d}u - \int_0^p \int_0^v f(u,v)\mathrm{d}u\mathrm{d}v.$$

对上式右端的两个积分分别做$\{u=s, v=1-t(1-s)\}$和$\{u=st, v=t\}$变换，得到

$$P\{X_{(k)} < \xi_p < X_{(l)}\} = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)}\int_0^p s^{k-1}(1-s)^{n-k}\mathrm{d}s - \frac{\Gamma(n+1)}{\Gamma(l)\Gamma(n-l+1)}\int_0^p t^{l-1}(1-t)^{n-l}\mathrm{d}t. \tag{6.3.1}$$

我们看到区间$(X_{(k)}, X_{(l)})$盖住分位数ξ_p的概率与总体分布$F(x)$无关，其置信水平为两个不完全贝塔函数之差.

在实际工作中，通常给定置信水平$1-\alpha$，我们选择置信下限$X_{(k)}$，使

$$P\{\xi_p \leqslant X_{(k)}\} = P\{p \leqslant F(X_{(k)})\} \leqslant \frac{\alpha}{2}, \tag{6.3.2}$$

选择置信上限$X_{(l)}$，使

$$P\{\xi_p \geqslant X_{(l)}\} = P\{p \geqslant F(X_{(l)})\} \leqslant \frac{\alpha}{2}. \tag{6.3.3}$$

故选k使不完全贝塔函数

$$\frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)}\int_0^p t^{k-1}(1-t)^{n-k}\mathrm{d}t \geqslant 1-\frac{\alpha}{2}, \tag{6.3.4}$$

而选l使

$$\frac{\Gamma(n+1)}{\Gamma(l)\Gamma(n-l+1)}\int_0^p t^{l-1}(1-t)^{n-l}\mathrm{d}t \leqslant \frac{\alpha}{2}. \tag{6.3.5}$$

对于单侧置信区间$[X_{(k)}, +\infty)$或$(-\infty, X_{(l)}]$，k或l满足

$$\frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)}\int_0^p z^{k-1}(1-z)^{n-k}\mathrm{d}z \geqslant 1-\alpha, \tag{6.3.6}$$

$$\frac{\Gamma(n+1)}{\Gamma(l)\Gamma(n-l+1)}\int_0^p z^{l-1}(1-z)^{n-l}\mathrm{d}z \leqslant \alpha. \tag{6.3.7}$$

在大样本的情形，则可利用样本p分位数ξ_p的渐近正态性，构成

置信区间. $F(X_{(k)})$为区间(0,1)均匀分布样本的$\frac{k}{n}$分位数,故

$$\frac{\sqrt{n}\,(nF(X_{(k)})-k)}{\sqrt{k(n-k)}}\text{ 渐近正态 } N(0,1)\text{ 分布},$$

有
$$\begin{aligned}P\{\xi_p \leqslant X_{(k)}\} &= P\{p \leqslant F(X_{(k)})\}\\ &= P\left\{\frac{\sqrt{n}\,(nF(X_{(k)})-k)}{\sqrt{k(n-k)}} \geqslant \frac{\sqrt{n}\,(np-k)}{\sqrt{k(n-k)}}\right\}\\ &\approx \Phi\left(-\frac{\sqrt{n}\,(np-k)}{\sqrt{k(n-k)}}\right).\end{aligned}$$

从而 k 应满足

$$-\frac{\sqrt{n}\,(np-k)}{\sqrt{k(n-k)}} \leqslant u, \tag{6.3.8}$$

其中 u 为由置信水平 $1-\alpha$ 确定的标准正态分布的分位数. 满足上式的最大整数 k 即为所求.

类似地

$$\begin{aligned}P\{\xi_p \geqslant X_{(l)}\} &= P\{p \geqslant F(X_{(l)})\}\\ &= P\left\{\frac{\sqrt{n}\,(nF(X_{(l)})-l)}{\sqrt{l(n-l)}} \leqslant \frac{\sqrt{n}\,(np-l)}{\sqrt{l(n-l)}}\right\}\\ &\approx \Phi\left(\frac{\sqrt{n}\,(np-l)}{\sqrt{l(n-l)}}\right),\end{aligned}$$

从而 l 应满足

$$\frac{\sqrt{n}\,(np-l)}{\sqrt{l(n-l)}} \leqslant u, \tag{6.3.9}$$

满足上式的最小整数 l 即为所求. 通常有

$$l = n-k+1.$$

对于中位数 $\xi_{1/2}$,求 k,l 有一个计算简单的近似公式:

$$\begin{aligned}k &= \left[\frac{1}{2}(n+1-u\sqrt{n+0.5-0.25u^2})\right],\\ l &= n-k+1.\end{aligned} \tag{6.3.10}$$

上式中[·]表示取整.

例 1 对 34 只晶体管做寿命试验，寿命的记录如下(单位：周)

寿长	3	4	5	6	7	8	9	10	11
频数	1	1	1	2	1	2	3	2	3

寿长	13	17	19	25	29	33	42	≥52
频数	5	2	2	1	1	1	2	4

晶体管寿命分布的中位数的 0.95 置信水平的单侧区间估计为$[X_{(k)},+\infty)$，由近似公式

$$\begin{aligned}&\frac{1}{2}(n+1-u\sqrt{n+0.5-0.25u^2})\\&=\frac{1}{2}[34+1-(1.645)\sqrt{34+0.5-0.25(1.645)^2}]\\&=12.72,\end{aligned}$$

故 $k=12$，$X_{(k)}=X_{(12)}=10$(周). 故分布中位数的 0.95 置信水平的单侧置信区间为$[10,+\infty)$.

考虑双侧区间估计：

$$\begin{aligned}&\frac{1}{2}(n+1-u\sqrt{n+0.5-0.25u^2})\\&=\frac{1}{2}[34+1-(1.960)\sqrt{34+0.5-0.25(1.960)^2}]\\&=11.82,\end{aligned}$$

故 $k=11$，$l=n-k+1=24$，中位数的 0.95 置信水平的双侧置信区间为$[X_{(k)},X_{(l)}]=[9,19]$.

§6.4 随机变量的容忍区间

容忍区间与置信区间是两个十分相似的概念.

定义 4.1 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，是来自分

布 $F(x)$ 的样本，$T_1(X_1,\cdots,X_n)$，$T_2(X_1,\cdots,X_n)$ 是两个统计量. 对任一 $(X_1,\cdots,X_n)$，$T_1\leqslant T_2$，若满足

$$P\{F(T_2(X_1,\cdots,X_n))-F(T_1(X_1,\cdots,X_n))\geqslant\beta\}\geqslant\gamma,\tag{6.4.1}$$

其中 $0<\beta<1$，$0<\gamma<1$，则称区间 $[T_1,T_2]$ 为总体 X 的 γ 概率水平 β 容忍量的**容忍区间**. T_1 称**容忍下限**，T_2 称**容忍上限**.

上述定义也可写成

$$P\{P[T_1<X\leqslant T_2]\geqslant\beta\}\geqslant\gamma,\tag{6.4.2}$$

这就是说容忍区间 $[T_1,T_2]$ 在 γ 水平上包容了总体 X 取值的 β 部分.

由样本次序统计量可构成分布任意的非参数容忍区间.

若 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 是一连续总体分布 $F(x)$ 的样本次序统计量，则总体分布包容在两个次序变量区间 $[X_{(k_1)},X_{(k_1+k_2)}]$ 的概率

$$U_{k_2}=F(X_{(k_1+k_2)})-F(X_{(k_1)})\tag{6.4.3}$$

是一个随机变量. $(F(X_{(k_1)}),F(X_{(k_1+k_2)}))$ 的二维联合分布为二变量 Dirichlet 分布，密度函数为

$$f(u,v)=\frac{\Gamma(n+1)}{\Gamma(k_1)\Gamma(k_2)\Gamma(n-k_1-k_2+1)}\cdot u^{k_1-1}(v-u)^{k_2-1}(1-v)^{n-k_1-k_2}.$$

由此可算出 $U_{k_2}=F(X_{(k_1+k_2)})-F(X_{(k_1)})$ 的分布为贝塔分布 $\mathrm{B}(k_2,n-k_2+1)$. 它与原总体分布函数 $F(x)$ 无关. 对给定的 β，概率

$$P\{U_{k_2}\geqslant\beta\}=\gamma$$

可由 $\mathrm{B}(k_2,n-k_2+1)$ 算出，其值与 k_2 有关，是一个不完全贝塔积分

$$P\{U_{k_2}\geqslant\beta\}=\int_\beta^1\frac{\Gamma(n+1)}{\Gamma(k_2)\Gamma(n-k_2+1)}z^{k_2-1}(1-z)^{n-k_2+1}\mathrm{d}z$$

$$\stackrel{\text{def}}{=\!=}\mathrm{B}_\beta(k_2,n-k_2+1).\tag{6.4.4}$$

当我们选取对称的次序统计量区间，即 $k_1=k$，$k_1+k_2=n-k+1$

时，$k_2=n-2k+1$. 寻求 γ 水平的 β 容忍量的容忍区间，即求 k，使不完全贝塔积分 $B_\beta(n-2k+1,2k)\geqslant\gamma$，$k$ 选择尽可能大的值，则以 $[X_{(k)},X_{(n-k+1)}]$ 作为总体 $F(x)$ 的容忍区间.

当然对给定的 $0<\beta<1$，$0<\gamma<1$，当 n 较小时，所求的容忍区间不一定存在，但由贝塔分布的性质知道，当 k 固定而 n 充分大时，$B_\beta(n-2k+1,2k)$ 可任意接近 1. 因而加大样本量则可获得要求的容忍区间.

例如当我们给定 $\beta=0.99$，$\gamma=0.95$ 时，需要 $n=473$，才能使 $[X_{(1)},X_{(n)}]$ 成为 γ 水平 β 容忍量的容忍区间.

类似于区间估计，也可考虑单侧容忍限的容忍区间.

§6.5 分布函数的置信区间

当对总体分布类型一无所知时，通常以样本经验分布函数作为总体分布的估计. 确定这样的估计的偏差是困难的，柯尔莫哥洛夫在 1933 年给出了这一问题的渐近解，这就是著名的关于经验分布的柯氏定理.

定理 5.1 设总体分布 $F(x)$ 连续，从总体中取一样本量为 n 的简单随机样本，所构成的样本经验分布函数记为 $F_n(x)$，令

$$D_n=\sup_x|F_n(x)-F(x)|, \tag{6.5.1}$$

则

$$\lim_{n\to\infty}P\left\{D_n\geqslant\frac{\lambda}{\sqrt{n}}\right\}=1-2\sum_{i=1}^{\infty}(-1)^{i+1}\mathrm{e}^{-2i^2\lambda^2},\quad \lambda>0. \tag{6.5.2}$$

柯尔莫哥洛夫给出的证明非常复杂. 1948 年 W. Feller 在论文(On the Kolmogorov-Smirnov limit theorems for empirical distributions, *Ann. Math. Stat.*, Vol 19, 177～189)中给出了一个简单一点的证明，有兴趣的读者可参阅该论文. 对于 D_n 统计量，容易看到它是一个分布任意的非参数统计量. 这是因为 D_n 必

在 $x=X_1,\cdots,X_n$ 这 n 个样本值点上达到. 而 $F(x)$ 连续时, $F(X_1)$, $\cdots$, $F(X_n)$ 可看成是 $(0,1)$ 上均匀分布的独立同分布样本, 由它们构成的经验分布记为 $F_n^*(x)$, $(0,1)$ 均匀分布记为 $F^*(x)$, 则

$$D_n^* = \sup_x |F_n^*(x) - F^*(x)| = D_n.$$

故可看出统计量 D_n 的分布与原总体分布 $F(x)$ 无关.

附表 12 对不同的 λ 值给出了这一极限分布的函数值, 当 n 较大时, 可利用它构成分布函数的置信区间. 即给定置信水平 $1-\alpha$, 选取 λ 的值, 使

$$Q(\lambda) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i+1}e^{-2i^2\lambda^2} = 1 - \alpha. \quad (6.5.3)$$

当 n 较大时, 近似有

$$P\left\{D_n \leqslant \frac{\lambda}{\sqrt{n}}\right\} = P\left\{\sup_x |F_n(x) - F(x)| \leqslant \frac{\lambda}{\sqrt{n}}\right\} = 1 - \alpha.$$

因而对任意 x, $F(x)$ 的 $1-\alpha$ 的置信区间为

$$\left[F_n(x) - \frac{\lambda}{\sqrt{n}},\ F_n(x) + \frac{\lambda}{\sqrt{n}}\right].$$

例 1 一工厂生产一种钢, 由于炼钢中的各种偶然因素的影响, 各炉钢的含硅量有差异, 因而把该厂成品钢中的含硅量看成一个随机变量. 现在采集了 120 个样本单元, 得到样本经验分布函数 $F_n(x)$. 由表 12 可以查出, 置信水平取为 $1-\alpha=0.95$ 时, $\lambda=1.36$, 故有

$$P\left\{\sup_x |F_n(x) - F(x)| \leqslant \frac{1.36}{\sqrt{120}}\right\} \approx 0.95,$$

从而对一切 x, $F(x)$ 的 0.95 水平的置信区间为

$$[F_n(x) - 0.124,\ F_n(x) + 0.124].$$

定理 5.1 的柯氏定理更经常用于总体分布拟合检验. 对原假设 H_0: $F(x)=F_0(x)$, 即检验总体分布是否为一指定的分布 $F_0(x)$, 可利用 $D_n=\sup_x |F_n(x)-F_0(x)|$ 作为检验统计量. 当 H_0 成立时, 柯氏定理给出了 D_n 的近似分布, 从而根据给定的显著性

水平 α，可查表找出临界值，确定出否定域. 对较小的 n 值可以从附表 13 查出 D_n 的临界值.

§6.6 Exceedence 统计量

这一节介绍另外一类用到经验分布函数的统计量，它们被称为 Exceedence 统计量.

由总体的一个简单随机样本 $X_1,\cdots,X_m$ 可构成一个经验分布函数 $F_m(x)$，而 $Y_1,\cdots,Y_n$ 是另一个简单随机样本，则 $F_m(Y_1),\cdots,F_m(Y_n)$ 是由样本 $X_1,\cdots,X_m,Y_1,\cdots,Y_n$ 构成的统计量，称为 **Exceedence 统计量**.

定理 6.1 若总体分布有密度函数 $f(t)$，$F_m(Y_j)$ $(j=1,\cdots,n)$ 的定义如上，记 $S_j=F_m(Y_j)$，则 $S_j(j=1,\cdots,n)$ 均在点集 $\left\{0,\frac{1}{m},\cdots,\frac{m-1}{m},1\right\}$ 上均匀分布.

证明 S_j 只在 $\left\{0,\frac{1}{m},\cdots,1\right\}$ 上取值是显然的.

约定 $X_{(0)}=-\infty$，$X_{(m+1)}=+\infty$. 对 $i=0,1,\cdots,m$ 和 $j=1,\cdots,n$，有

$$
\begin{aligned}
P\left\{S_j=\frac{i}{m}\right\}&=P\{X_{(i)}\leqslant Y_j\leqslant X_{(i+1)}\}\\
&=\int_{-\infty}^{+\infty}P\{X_{(i)}\leqslant Y_j\leqslant X_{(i+1)}|Y_j=t\}f(t)\mathrm{d}t\\
&=\int_{-\infty}^{+\infty}P\{X_{(i)}\leqslant t\leqslant X_{(i+1)}\}f(t)\mathrm{d}t\\
&=\int_{-\infty}^{+\infty}\binom{m}{i}[F(t)]^i[1-F(t)]^{m-i}f(t)\mathrm{d}t.
\end{aligned}
$$

上面的第三个等号是由于 Y_j 与 $X_1,\cdots,X_m$ 独立. 对最后的积分再做变换 $F(t)=u$，则

$$
P\left\{S_j=\frac{i}{m}\right\}=\int_0^1\binom{m}{i}u^i(1-u)^{m-i}\mathrm{d}u
$$

$$= \frac{1}{m+1}\left[\frac{(m+1)!}{i!(m-i)!}\int_0^1 u^i(1-u)^{m-i}\mathrm{d}u\right]$$

$$= \frac{1}{m+1}. \tag{6.6.1}$$

证毕.

定理 6.2 条件同定理 6.1,$S_{(1)},\cdots,S_{(n)}$为$S_1,\cdots,S_n$的次序统计量,则有

$$P\left\{S_{(j)}=\frac{i}{m}\right\}=\begin{cases}\dfrac{\dbinom{m+n-i-j}{m-i}\dbinom{i+j-1}{i}}{\dbinom{m+n}{m}}, & \text{当 } i=0,1,\cdots,m \text{ 时},\\ 0, & \text{其他}.\end{cases}$$

证明 由$S_{(j)}$定义有$S_{(j)}=F_m(Y_{(j)})$ $(j=1,\cdots,n)$,对$i=0,1,\cdots,m$,

$$P\left\{S_{(j)}=\frac{i}{m}\right\}=P\{X_{(i)}\leqslant Y_{(j)}<X_{(i+1)}\}.$$

若将混合样本$X_1,\cdots,X_m,Y_1,\cdots,Y_n$对应的秩统计量记为$Q_1,\cdots,Q_m,R_1,\cdots,R_n$,则事件

$$\begin{aligned}\{X_{(i)}\leqslant Y_{(j)}<X_{(i+1)}\}\\ =\{&(1,\cdots,i+j-1)\text{ 中有 } i \text{ 个 } Q,\\ &(i+j+1,\cdots,m+n)\text{ 中有 } m-i \text{ 个 } Q\}.\end{aligned}$$

由于$(Q_1,\cdots,Q_m,R_1,\cdots,R_n)$在$(1,\cdots,m+n)$的全部排列上均匀分布,所以

$$P\left\{S_{(j)}=\frac{i}{m}\right\}=\frac{\dbinom{i+j-1}{i}\dbinom{m+n-i-j}{m-i}}{\dbinom{m+n}{m}},$$

$$i=0,1,\cdots,m, \tag{6.6.2}$$

其他情形为 0 是显然的. 证毕.

由上述定理 6.1 和定理 6.2 知$(S_1,\cdots,S_n)$和$(S_{(1)},\cdots,S_{(n)})$均

是分布任意的非参数统计量. 但应注意 $S_1,\cdots,S_n$ 并不相互独立.

例 1 预测问题. 从一连续分布总体中抽取一个简单随机样本 $X_1,\cdots,X_m$, 若再取一个简单随机样本 $Y_1,\cdots,Y_n$, 当 n 为奇数时, 求样本中位数 $Y_{\left(\frac{n+1}{2}\right)}$ 的预测区间.

由定理 6.2 知, 对 $1\leqslant r_1<r_2\leqslant m$, 有

$$P\{X_{(r_1)}\leqslant Y_{\left(\frac{n+1}{2}\right)}\leqslant X_{(r_2)}\}$$

$$=P\left\{\frac{r_1}{m}\leqslant F_m(Y_{\left(\frac{n+1}{2}\right)})\leqslant\frac{r_2-1}{m}\right\}$$

$$=\sum_{i=r_1}^{r_2-1}\left[\binom{m+\frac{n-1}{2}-i}{m-i}\binom{\frac{n-1}{2}+i}{i}\Big/\binom{m+n}{m}\right].$$

对给定的信赖水平 $1-\alpha$, 可求 r_1,r_2 满足

$$\sum_{i=0}^{r_1-1}\left[\binom{m+\frac{n-1}{2}-i}{m-i}\binom{\frac{n-1}{2}+i}{i}\Big/\binom{m+n}{m}\right]\leqslant\frac{\alpha}{2}.$$

$$\sum_{i=r_2}^{m}\left[\binom{m+\frac{n-1}{2}-i}{m-i}\binom{\frac{n-1}{2}+i}{i}\Big/\binom{m+n}{m}\right]\leqslant\frac{\alpha}{2},$$

则 $[X_{(r_1)},X_{(r_2)}]$ 即为 $Y_{\left(\frac{n+1}{2}\right)}$ 的预测区间.

最后, 我们指出第一章中的 Mann-Whitney 统计量可表成 Exceedence 统计量的形式:

$$U=\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(Y_j-X_i)$$

$$=\sum_{j=1}^{n}\#(\text{小于 } Y_j \text{ 的 } X_i \text{ 的个数})=m\sum_{j=1}^{n}S_j. \quad (6.6.3)$$

习 题

1. 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布, 是来自总体分布 $F(x)$ (密度函数 $f(x)$) 的样本, 样本次序统计量记为 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$. 求样本差程 $R=X_{(n)}-X_{(1)}$ 的分布密度.

2. 假设条件同第 1 题，$\xi_{1/2}$是总体中位数，证明

$$P\{X_{(k)} < \xi_{1/2} < X_{(n-k+1)}\} = 1 - 2\int_0^{\frac{1}{2}} \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} z^{n-k}(1-z)^{k-1}\mathrm{d}z.$$

3. 设总体分布密度函数

$$f(x) = \begin{cases} \mathrm{e}^{-x}, & x > 0, \\ 0, & x \leqslant 0, \end{cases}$$

求当 $n \to \infty$时，容量为 n 的简单随机样本的样本中位数的极限分布.

4. 参看 §6.3 中晶体管寿命的例子，对晶体管寿命分布的$\frac{1}{4}$分位数 $\xi_{1/4}$作点估计和 0.95 置信水平的单侧下限区间估计.

5. 对一批电器元件，抽取 24 个做加速寿命试验，测得 24 个元件的寿命数据为(单位：小时)：

575，	778，	880，	969，	984，	1003，
1008，	1021，	1031，	1034，	1053，	1054，
1226，	1393，	1439，	1480，	1513，	1611，
1612，	1612，	1624，	1627，	1631，	1768，

求这批元件寿命分布的中位数的 0.95 置信水平的区间估计.

6. 设 $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$是来自连续总体分布 $F(x)$的样本次序统计量. 证明：对 $0<\beta<1$，

$$P\{F(X_{(n)}) - F(X_{(1)}) > \beta\} = 1 - n\beta^{n-1} + (n-1)\beta^n.$$

7. 设 $F_m(x)$是总体 $F(x)$的一个样本量为 m 的简单随机样本构成的经验分布函数. 对任意两个实数 $x<y$，求 $\mathrm{cov}(F_m(x), F_m(y))$.

8. 设 S_j $(j=1,\cdots,n)$是 Exceedence 统计量(如 §6.6 中定义)，求 S_k 和 S_l $(1\leqslant k<l\leqslant n)$的联合分布.

第七章　Hodges-Lehmann 估计

§7.1　Hodges-Lehmann 估计量

我们在数理统计课程中学过的许多点估计方法其实是一些非参数点估计方法，如矩估计法、最小二乘估计法、最小一乘估计法等等，这些方法确定估计量的准则均与总体分布无关. 在这一章中介绍一种与前面几章所述检验问题关系密切的 Hodges-Lehmann 估计. 以下简称 H-L 估计.

首先介绍一样本位置参数的 H-L 估计技术.

设随机变量 $X_1, \cdots, X_n$ 相互独立同分布，分布为 $F(x-\theta)$，分布函数 $F(t)$ 连续，关于 0 点对称，θ 为位置参数. 若对假设检验问题

$$H_0: \theta = 0, \qquad H_1: \theta > 0$$

有检验统计量 $V(X_1, \cdots, X_n)$ 满足下列三个条件：

(1) 检验的否定域为 $\{V(\cdot) \geqslant c\}$，c 为临界值；

(2) 对任一指定的 $(x_1, \cdots, x_n)$，$V(x_1+h, \cdots, x_n+h)$ 是 h 的非降函数；

(3) 当 H_0: $\theta=0$ 成立时，对任一满足假定的分布 $F(x)$，$V(X_1, \cdots, X_n)$ 的分布关于某一 ξ 对称，

则可利用统计量 $V(X_1, \cdots, X_n)$ 构造一个位置参数 θ 的 H-L 估计.

例 1　当总体分布为 $F(x-\theta)$，函数 $F(t)$ 关于 0 点对称时，符号统计量、符号秩统计量、T 统计量，即

$$B = \sum_{i=1}^{n} \Psi_i, \ \Psi_i = 1, \text{当 } X_i > 0, \text{或 } 0, \text{当 } X_i \leqslant 0;$$

$$W^+=\sum_{i=1}^{n}\Psi_i R_i^+,\ (R_1^+,\cdots,R_n^+)\text{为}(X_1,\cdots,X_n)\text{的绝对秩};$$

$$T=\frac{\overline{X}}{S\sqrt{n}},\ \overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\ S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

均满足上述条件(1)～(3). 首先这几个统计量作检验的否定域均具有条件(1)要求的形式. 统计量

$$B=\sum_{i=1}^{n}\Psi_i=\sum_{i=1}^{n}\psi(X_i),$$

故
$$B(x_1+h,\cdots,x_n+h)=\sum_{i=1}^{n}\psi(x_i+h)$$

是 h 的非降函数. 统计量

$$W^+=\sum_{i=1}^{n}\Psi_i R_i^+=\sum_{i=1}^{n}\sum_{j\geqslant i}\psi(X_i+X_j),$$

故
$$W^+(x_1+h,\cdots,x_n+h)=\sum_{i=1}^{n}\sum_{j\geqslant i}\psi(x_i+x_j+2h)$$

是 h 的非降函数. 统计量 T 有

$$T(x_1+h,\cdots,x_n+h)=\frac{\overline{x}}{S/\sqrt{n}}+\frac{h}{S/\sqrt{n}}$$

是 h 的非降函数. 当总体分布 $F(x)$ 关于 0 点对称时, B 的分布为二项分布 $b\left(n,\frac{1}{2}\right)$, 有对称点 $\frac{n}{2}$. W^+ 的分布也是对称的, 对称点为 $\frac{n(n+1)}{4}$(参见第一章定理 4.5). 对于统计量 T, 由于

$$(X_1,\cdots,X_n)\overset{d}{=\!=}(-X_1,\cdots,-X_n),$$

故
$$T=\frac{\overline{X}}{S/\sqrt{n}}\overset{d}{=\!=}\frac{-\overline{X}}{S/\sqrt{n}}=-T.$$

T 的分布关于 0 点对称.

定义 1.1 统计量 $V(X_1,\cdots,X_n)$ 如上所述, 则基于 $V(\cdot)$ 的 θ 的 H-L 估计量为

$$\hat{\theta}=\frac{\theta^*+\theta^{**}}{2},\tag{7.1.1}$$

其中

$$\theta^{*}(X_1,\cdots,X_n)=\sup\{\theta|V(X_1-\theta,\cdots,X_n-\theta)>\xi\},$$

$$\theta^{**}(X_1,\cdots,X_n)=\inf\{\theta|V(X_1-\theta,\cdots,X_n-\theta)<\xi\}.$$

当 $V(X_1-\theta,\cdots,X_n-\theta)$ 是 θ 的连续、严格单调函数时，则有 $\hat{\theta}=\theta^{*}=\theta^{**}$，有 $V(X_1-\hat{\theta},\cdots,X_n-\hat{\theta})=\xi$. 但一般情况下，有 $\theta^{**}>\theta^{*}$.

例 2 由统计量 $T=\dfrac{\overline{X}}{S/\sqrt{n}}$ 产生 θ 的 H-L 估计. 此时，

$$T(X_1-\theta,\cdots,X_n-\theta)=\frac{\overline{X}-\theta}{S/\sqrt{n}}$$

是 θ 的连续严格单调降函数. 所以

$$\hat{\theta}=\theta^{*}=\theta^{**},\quad \frac{\overline{X}-\hat{\theta}}{S\sqrt{n}}=\xi=0,\quad \hat{\theta}=\overline{X}.$$

由符号秩统计量 $W^{+}=\sum_{i=1}^{n}\Psi_i R_i^{+}=\sum_{i=1}^{n}\sum_{j\geqslant i}\psi(X_i+X_j)$ 产生 θ 的 H-L 估计，此时

$$\begin{aligned}W^{+}(X_1-\theta,\cdots,X_n-\theta)&=\sum_{i=1}^{n}\sum_{j\geqslant i}\psi(X_i+X_j-2\theta)\\&=\sum_{i=1}^{n}\sum_{j\geqslant i}\psi\left(\frac{X_i+X_j}{2}-\theta\right).\end{aligned}$$

以 $W_{(1)}\leqslant\cdots\leqslant W_{(M)}$ $(M=n(n+1)/2)$ 记 M 个随机变量

$$\left\{\frac{X_i+X_j}{2},\quad i\leqslant j,i,j=1,\cdots,n\right\}$$

的次序统计量. 若 M 是奇数，则 $\xi=\dfrac{n(n+1)}{4}=\dfrac{M}{2}$ 可表示为 $\xi=K+\dfrac{1}{2}$，K 为正整数. 由此得

$$\begin{aligned}\theta^{*}&=\sup\left\{\theta\left|W^{+}(X_1-\theta,\cdots,X_n-\theta)>K+\frac{1}{2}\right.\right\}\\&=\sup\left\{\theta\left|\sum_{i=1}^{n}\sum_{j\geqslant i}\psi\left(\frac{X_i+X_j}{2}-\theta\right)\geqslant K+1\right.\right\}\end{aligned}$$

$$= \sup\{\theta \mid \psi(W_{(K+1)} - \theta) = 1\}$$

$$= \sup\{\theta \mid (W_{(K+1)} - \theta) > 0\} = W_{(K+1)},$$

类似地可求得

$$\theta^{**} = \inf\left\{\theta \,\middle|\, \sum_{i=1}^{n}\sum_{j\geqslant i}\psi\left(\frac{X_i + X_j}{2} - \theta\right) < K + \frac{1}{2}\right\}$$

$$= \inf\{\theta \mid \psi(W_{(K+1)} - \theta) = 0\}$$

$$= \inf\{\theta \mid (W_{(K+1)} - \theta) \leqslant 0\} = W_{(K+1)}.$$

所以

$$\hat{\theta} = \frac{\theta^* + \theta^{**}}{2} = W_{(K+1)}.$$

若 M 是偶数，则 $M=2K$，$\xi=\frac{M}{2}=K$，K 为正整数. 由此得

$$\theta^* = \sup\{\theta \mid W^+(X_1 - \theta, \cdots, X_n - \theta) > K\}$$

$$= \sup\left\{\theta \,\middle|\, \sum_{i=1}^{n}\sum_{j\geqslant i}\psi\left(\frac{X_i + X_j}{2} - \theta\right) \geqslant K + 1\right\}$$

$$= \sup\{\theta \mid \psi(W_{(K)} - \theta) = 1\}$$

$$= \sup\{\theta \mid (W_{(K)} - \theta) > 0\} = W_{(K)},$$

$$\theta^{**} = \inf\left\{\theta \,\middle|\, \sum_{i=1}^{n}\sum_{j\geqslant i}\psi\left(\frac{X_i + X_j}{2} - \theta\right) < K\right\}$$

$$= \inf\{\theta \mid \psi(W_{(K+1)} - \theta) = 0\}$$

$$= \inf\{\theta \mid (W_{(K+1)} - \theta) \leqslant 0\} = W_{(K+1)},$$

$$\hat{\theta} = \frac{\theta^* + \theta^{**}}{2} = \frac{W_{(K)} + W_{(K+1)}}{2}.$$

总之，$\hat{\theta}$ 是 $\left\{\frac{X_i+X_j}{2},\ i\leqslant j,\ i,j=1,\cdots,n\right\}$ 的中位数.

由符号统计量 $B=\sum_{i=1}^{n}\Psi_i$ 产生 θ 的 H-L 估计，此时

$$B(X_1 - \theta, \cdots, X_n - \theta) = \sum_{i=1}^{n}\psi(X_i - \theta).$$

以 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 记样本 $(X_1,\cdots,X_n)$ 的次序统计量. 与上面求 W^+ 的对应的 H-L 估计类似，可以求得：当 n 为奇数，$n=2K+1$

时，$\hat{\theta}=X_{(K+1)}$；当 n 为偶数，$n=2K$ 时，$\hat{\theta}=\frac{X_{(K)}+X_{(K+1)}}{2}$. $\hat{\theta}$ 是 $\{X_i,i=1,\cdots,n\}$的中位数.

对二样本位置参数可以类似地定义 H-L 估计量：样本 $X_1,\cdots,X_m$ 相互独立同分布，分布为 $F(x)$；$Y_1,\cdots,Y_n$ 相互独立同分布，分布为 $F(x-\Delta)$，$F(x)$连续.若假设检验问题

$$H_0: \Delta=0, \qquad H_1: \Delta>0$$

有检验统计量 $V(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$满足下列条件：

(1) 否定域有形式$\{V(\cdot\,;\cdot)\geqslant c\}$，$c$ 为临界值；

(2) 对任一$(x_1,\cdots,x_m;y_1,\cdots,y_n)$，$V(x_1,\cdots,x_m;\ y_1+h,\cdots,y_n+h)$是 h 的非降函数；

(3) 在 H_0 下，对任一满足假定的总体分布 $F(x)$，$V(\cdot\,;\cdot)$的分布关于某个 ξ 对称.

定义 1.2 统计量$V(\cdot\,;\cdot)$如上所述，基于 $V(\cdot\,;\cdot)$的参数 Δ 的 H-L 估计量是

$$\hat{\Delta}=\frac{\Delta^{*}+\Delta^{**}}{2}, \tag{7.1.2}$$

其中

$$\begin{aligned}&\Delta^{*}(X_1,\cdots,X_m;Y_1,\cdots,Y_n)\\&\quad=\sup\{\Delta|V(X_1,\cdots,X_m;Y_1-\Delta,\cdots,Y_n-\Delta)>\xi\},\\&\Delta^{**}(X_1,\cdots,X_m;Y_1,\cdots,Y_n)\\&\quad=\inf\{\Delta|V(X_1,\cdots,X_m;Y_1-\Delta,\cdots,Y_n-\Delta)<\xi\}.\end{aligned}$$

例 3 考虑 Wilcoxon 秩和统计量

$$W=\sum_{i=1}^{n}R_i=\frac{n(n+1)}{2}+\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(Y_j-X_i)$$

(参见第一章第 3 节). 它满足上述条件(1)～(3)，可以求得由 W 产生的参数 Δ 的 H-L 估计量 $\hat{\Delta}$ 是$\{Y_j-X_i,\ i=1,\cdots,m,\ j=1,\cdots,n\}$的中位数.

§ 7.2 Hodges-Lehmann 估计量的小样本性质

设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布为 $F(x-\theta)$，函数 $F(t)$ 连续.

引理 2.1 设一样本位置参数检验问题统计量 $V(X_1,\cdots,X_n)$ 满足条件(1)～(3)，$\hat{\theta}(X_1,\cdots,X_n)$ 是基于 $V(\cdot)$ 的 θ 的 H-L 估计量，则 $\hat{\theta}(X_1,\cdots,X_n)$ 是平移的（或称位置同变的），即对任意一个 $(X_1,\cdots,X_n)$ 及 K，有

$$\hat{\theta}(X_1+K,\cdots,X_n+K)=\hat{\theta}(X_1,\cdots,X_n)+K.$$

证明 H-L 估计定义中的统计量

$$\begin{aligned}
&\theta^*(X_1+K,\cdots,X_n+K)\\
&\quad=\sup\{\theta|V(X_1+K-\theta,\cdots,X_n+K-\theta)>\xi\}\\
&\quad=\sup\{\theta|V(X_1-(\theta-K),\cdots,X_n-(\theta-K))>\xi\}\\
&\quad=\sup\{\tau+K|V(X_1-\tau,\cdots,X_n-\tau)>\xi\}\\
&\quad=K+\sup\{\tau|V(X_1-\tau,\cdots,X_n-\tau)>\xi\}\\
&\quad=K+\theta^*(X_1,\cdots,X_n),
\end{aligned}$$

类似地有

$$\theta^{**}(X_1+K,\cdots,X_n+K)=K+\theta^{**}(X_1,\cdots,X_n).$$

所以

$$\hat{\theta}=\frac{\theta^*+\theta^{**}}{2}$$

是平移的. 证毕.

引理 2.2 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布函数为 $F(x-\theta)$，又 $T(X_1,\cdots,X_n)$ 是一平移的统计量，则对一切 t，有等式

$$P_\theta\{T(X_1,\cdots,X_n)-\theta\leqslant t\}=P_0\{T(X_1,\cdots,X_n)\leqslant t\},$$

其中 P_θ 表示 $X_i\sim F(x-\theta)$ $(i=1,\cdots,n)$ 时的概率，P_0 表示 $X_i\sim F(x)$ $(i=1,\cdots,n)$ 时的概率.

证明 可直接计算出概率：

$$
\begin{aligned}
&P_\theta\{T(X_1,\cdots,X_n)-\theta\leqslant t\}\\
&\qquad=P_0\{T(X_1-\theta,\cdots,X_n-\theta)\leqslant t\},\\
&\qquad=P_0\{T(X_1,\cdots,X_n)\leqslant t\}.
\end{aligned}
$$

证毕.

根据上述两条引理，在讨论 H-L 估计量的分布性质时，不失一般性，可假定 $\theta=0$.

定理 2.3 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布为 $F(x-\theta)$，函数 $F(t)$ 连续且关于 0 点对称，$\hat{\theta}(X_1,\cdots,X_n)$ 是基于检验统计量 $V(X_1,\cdots,X_n)$ 的 θ 的 H-L 估计量，$V(\cdot)$ 满足要求的条件(1)～(3)，且对 $(x_1,\cdots,x_n)$ 几乎处处有

$$
V(x_1,\cdots,x_n)+V(-x_1,\cdots,-x_n)=2\xi, \qquad (7.2.1)
$$

其中 ξ 是 $V(\cdot)$ 的分布的对称点，则 $\hat{\theta}(X_1,\cdots,X_n)$ 是奇统计量，即有

$$
\hat{\theta}(X_1,\cdots,X_n)=-\hat{\theta}(-X_1,\cdots,-X_n),
$$

并且 $\hat{\theta}(X_1,\cdots,X_n)$ 的分布关于 θ 对称.

证明 统计量

$$
\begin{aligned}
&\theta^*(-X_1,\cdots,-X_n)\\
&\qquad=\sup\{\theta|V(-X_1-\theta,\cdots,-X_n-\theta)>\xi\}\\
&\qquad=\sup\{\theta|2\xi-V(X_1+\theta,\cdots,X_n+\theta)>\xi\}\\
&\qquad=\sup\{-\theta|V(X_1-\theta,\cdots,X_n-\theta)<\xi\}\\
&\qquad=-\inf\{\theta|V(X_1-\theta,\cdots,X_n-\theta)<\xi\}\\
&\qquad=-\theta^{**}(X_1,\cdots,X_n).
\end{aligned}
$$

同理可证

$$
\theta^{**}(-X_1,\cdots,-X_n)=-\theta^*(X_1,\cdots,X_n).
$$

所以 $$\hat{\theta}(-X_1,\cdots,-X_n)=-\hat{\theta}(X_1,\cdots,X_n).$$

由 $\hat{\theta}(\cdot)$ 是奇统计量，且是平移的，可推得 $\hat{\theta}(\cdot)$ 的分布关于 θ 对称. 其论证如下：

$$\hat{\theta}(X_1,\cdots,X_n)-\theta=\hat{\theta}(X_1-\theta,\cdots,X_n-\theta).$$

但 $X_i-\theta\ (i=1,\cdots,n)$的分布关于 0 点对称，故

$$\hat{\theta}(X_1,\cdots,X_n)-\theta \overset{d}{=\!=} \hat{\theta}(\theta-X_1,\cdots,\theta-X_n)$$
$$=\theta+\hat{\theta}(-X_1,\cdots,-X_n)=\theta-\hat{\theta}(X_1,\cdots,X_n),$$

即有 $\hat{\theta}(\cdot)$的分布关于 θ 对称. 证毕

例 1 考虑 $\hat{\theta}(\cdot)$是基于检验统计量 $W^+(\cdot)$的 θ 的 H-L 估计量. $W^+(\cdot)$满足要求的条件(1)～(3)，且

$$W^+(X_1,\cdots,X_n)=\sum_{i=1}^{n}\sum_{j\geqslant i}\psi(X_i+X_j),$$

$$W^+(-X_1,\cdots,-X_n)=\sum_{i=1}^{n}\sum_{j\geqslant i}\psi(-(X_i+X_j)).$$

当 $F(x)$连续时，$\{X_i+X_j=0\}$的概率为 0，所以几乎处处有

$$W^+(X_1,\cdots,X_n)+W^+(-X_1,\cdots,-X_n)=\frac{n(n+1)}{2},$$

满足定理的条件，故对应的 H-L 估计

$$\hat{\theta}(X_1,\cdots,X_n)=\left\{\frac{X_i+X_j}{2},\ i\leqslant j,i,j=1,\cdots,n\right\}\text{的中位数},$$

有关于 θ 对称的分布.

由定理 2.3 知在定理条件下，H-L 估计 $\hat{\theta}$ 是 θ 的无偏估计量. 另外，在一定条件下 $\hat{\theta}$ 可以是 θ 的中位数无偏的估计.

定义 2.1 $\hat{\eta}$ 是 η 的估计量，若对任意的 η，有

$$P_\eta\{\hat{\eta}\leqslant\eta\}=\frac{1}{2}, \tag{7.2.2}$$

则称 $\hat{\eta}$ 是 η 的中位数无偏的估计量.

引理 2.4 设 $\hat{\theta}(X_1,\cdots,X_n)$是基于满足条件的检验统计量 $V(X_1,\cdots,X_n)$的 H-L 估计量，若样本 $X_1,\cdots,X_n$ 的分布连续，则 θ^* 和 θ^{**} 的分布函数连续.

证明 要证对任一给定的 t，对统计量 θ^*，有

$$P_\theta\{\theta^*(X_1,\cdots,X_n)=t\}=0.$$

令 $T_i=X_i-X_1\quad(i=2,\cdots,n)$，则由

$$V(x_1,\cdots,x_n)=V(x_1,x_1+t_2,\cdots,x_1+t_n)$$

是 x_1 的非降函数，故存在

$$u(t_2,\cdots,t_n)=\inf\{x_1|V(x_1,x_1+t_2,\cdots,x_1+t_n)>\xi\}$$

满足：

$$x_1<u(t_2,\cdots,t_n)\text{时},\ V(x_1,x_1+t_2,\cdots,x_1+t_n)\leqslant\xi;$$

$$x_1>u(t_2,\cdots,t_n)\text{时},\ V(x_1,x_1+t_2,\cdots,x_1+t_n)>\xi.$$

则有统计量

$$\begin{aligned}\theta^*(X_1,\cdots,X_n)&=\theta^*(X_1,X_1+T_2\cdots,X_1+T_n)\\&=\sup\{\theta|V(X_1-\theta,X_1+T_2-\theta,\cdots,X_1+T_n-\theta)>\xi\}\\&=X_1-u(T_2,\cdots,T_n).\end{aligned}$$

故

$$\begin{aligned}&\{(x_1,t_2,\cdots,t_n)|\theta^*(x_1,x_1+t_2,\cdots,x_1+t_n)=t\}\\&\quad=\{(x_1,t_2,\cdots,t_n)|x_1=u(t_2,\cdots,t_n)+t\}\end{aligned}$$

是 X_1 与 $(T_2,\cdots,T_n)$ 乘积空间的可测集. 给定 $(t_2,\cdots,t_n)=(t_2^0,\cdots,t_n^0)$ 的切口为

$$\{x_1|x_1=u(t_2^0,\cdots,t_n^0)+t\},$$

而 X_1 有连续分布，故此切口的概率测度为 0. 利用 Fubini 定理知：X_1 与 $(T_2,\cdots,T_n)$ 乘积空间一可测子集概率测度为 0 的充要条件为对几乎每一 $(t_2,\cdots,t_n)$ 的 X_1 切口具有零概率测度. 故由对任一 $(t_2^0,\cdots,t_n^0)$，集合 $\{x_1|x_1=u(t_2^0,\cdots,t_n^0)+t\}$ 的概率均为 0，知集合 $\{\theta^*=t\}$ 的概率为 0，即 θ^* 的分布函数连续.

类似地可证 θ^{**} 的分布函数连续. 证毕.

引理 2.5 设 $\hat{\theta}(X_1,\cdots,X_n)$ 是基于满足条件的检验统计量 $V(X_1,\cdots,X_n)$ 的 H-L 估计量，样本 $X_1,\cdots,X_n$ 的分布连续，则对任一实数 a，有

$$\begin{aligned}P_\theta\{V(X_1-a,\cdots,X_n-a)<\xi\}&\leqslant P_\theta\{\hat{\theta}(X_1,\cdots,X_n)<a\}\\&=P_\theta\{V(X_1-a,\cdots,X_n-a)\leqslant\xi\}.\end{aligned}\qquad(7.2.3)$$

证明 由 θ^*,θ^{**} 的定义,有

$$\begin{aligned}&\theta^*(x_1,\cdots,x_n)>a\\&\quad\Longrightarrow V(x_1-a,\cdots,x_n-a)>\xi\\&\quad\Longrightarrow\theta^*(x_1,\cdots,x_n)\geqslant a\end{aligned}$$

及

$$\begin{aligned}&\theta^{**}(x_1,\cdots,x_n)<a\\&\quad\Longrightarrow V(x_1-a,\cdots,x_n-a)<\xi\\&\quad\Longrightarrow\theta^{**}(x_1,\cdots,x_n)\leqslant a.\end{aligned}$$

而由引理 2.4 知 θ^*,θ^{**} 分布函数连续,故有

$$\begin{aligned}P_\theta\{\theta^*(X_1,\cdots,X_n)<a\}&=P_\theta\{\theta^*(X_1,\cdots,X_n)\leqslant a\}\\&=P_\theta\{V(X_1-a,\cdots,X_n-a)\leqslant\xi\};\\P_\theta\{\theta^{**}(X_1,\cdots,X_n)<a\}&=P_\theta\{\theta^{**}(X_1,\cdots,X_n)\leqslant a\}\\&=P_\theta\{V(X_1-a,\cdots,X_n-a)<\xi\}.\end{aligned}$$

而 $\theta^{**}(x_1,\cdots,x_n)\geqslant\theta^*(x_1,\cdots,x_n)$,由此推知

$$\begin{aligned}\{\theta^{**}(X_1,\cdots,X_n)<a\}&\subset\{\hat\theta(X_1,\cdots,X_n)<a\}\\&\subset\{\theta^*(X_1,\cdots,X_n)<a\}.\end{aligned}$$

所以

$$\begin{aligned}P_\theta\{V(X_1-a,\cdots,X_n-a)<\xi\}&=P_\theta\{\theta^{**}(X_1,\cdots,X_n)<a\}\\&\leqslant P_\theta\{\hat\theta(X_1,\cdots,X_n)<a\}\leqslant P_\theta\{\theta^*(X_1,\cdots,X_n)<a\}\\&=P_\theta\{V(X_1-a,\cdots,X_n-a)\leqslant\xi\}.\end{aligned}$$

证毕.

引理 2.6 设 $\hat\theta(X_1,\cdots,X_n)$是基于满足条件的检验统计量 $V(X_1,\cdots,X_n)$的 H-L 估计,样本 $X_1,\cdots,X_n$ 分布连续. 若 $\theta=0$ 时, $V(X_1,\cdots,X_n)$的分布的对称点为 ξ,则有

$$\frac{1-\varepsilon}{2}\leqslant P_\theta\{\hat\theta(X_1,\cdots,X_n)\leqslant\theta\}\leqslant\frac{1+\varepsilon}{2},\qquad(7.2.4)$$

其中 $\varepsilon=P_0\{V(X_1,\cdots,X_n)=\xi\}$.

证明 由 $V(\cdot)$的分布关于 ξ 对称,及 ε 的定义,有

$$P_0\{V(X_1,\cdots,X_n)<\xi\}=P_0\{V(X_1,\cdots,X_n)>\xi\}=\frac{1-\varepsilon}{2},$$

$$P_0\{V(X_1,\cdots,X_n)\leqslant\xi\}=P_0\{V(X_1,\cdots,X_n)\geqslant\xi\}=\frac{1+\varepsilon}{2}.$$

由引理 2.5，有

$$\frac{1-\varepsilon}{2}=P_0\{V(X_1,\cdots,X_n)<\xi\}\leqslant P_0\{\hat{\theta}(X_1,\cdots,X_n)<0\}$$
$$\leqslant P_0\{V(X_1,\cdots,X_n)\leqslant\xi\}=\frac{1+\varepsilon}{2}.$$

而

$$P_\theta\{\hat{\theta}(X_1,\cdots,X_n)\leqslant\theta\}=P_\theta\{\hat{\theta}(X_1-\theta,\cdots,X_n-\theta)\leqslant 0\}$$
$$=P_0\{\hat{\theta}(X_1,\cdots,X_n)\leqslant 0\},$$

由引理 2.4 知 $\hat{\theta}(\cdot)$分布连续，故有本引理结论. 证毕.

定理 2.7 设 $\hat{\theta}(X_1,\cdots,X_n)$是基于满足条件的检验统计量 $V(X_1,\cdots,X_n)$的 H-L 估计，样本 $X_1,\cdots,X_n$ 分布连续，$\theta=0$ 时，$V(X_1,\cdots,X_n)$的分布的对称点为 ξ. 若

$$P_0\{V(X_1,\cdots,X_n)=\xi\}=0,$$

则 $\hat{\theta}(X_1,\cdots,X_n)$是参数 θ 的中位数无偏估计.

证明 由引理 2.6 立即得定理结论. 证毕.

例 2 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布为 $F(x-\theta)$，分布函数 $F(t)$连续，中位数为 θ，则由符号统计量 $B(X_1,\cdots,X_n)=\sum_{i=1}^{n}\psi(X_i)$产生的 θ 的 H-L 估计是样本中位数，而 $\theta=0$ 时，$B(\cdot)$的分布关于$\frac{n}{2}$对称，当 n 为奇数时，显然有

$$P_0\left\{B(X_1,\cdots,X_n)=\frac{n}{2}\right\}=0.$$

所以样本中位数是 θ 的中位数无偏估计.

对二样本位置参数问题有类似的结果.

引理 2.8 由满足二样本位置问题条件(1)～(3)的检验统计量 $V(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$产生的参数 Δ 的 H-L 估计量 $\hat{\Delta}(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$是平移的，即对一切$(x_1,\cdots,x_m;y_1,\cdots,y_n)$及 K，

有

$$\hat{\Delta}(x_1,\cdots,x_m;y_1+K,\cdots,y_n+K)$$
$$=\hat{\Delta}(x_1,\cdots,x_m;y_1,\cdots,y_n)+K.$$

引理 2.9 设随机变量 $X_1,\cdots,X_m$ 相互独立同分布，分布为 $F(x)$；随机变量 $Y_1,\cdots,Y_n$ 相互独立同分布，分布为 $F(x-\Delta)$. 若统计量 $T(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$是平移的，则对一切 t，有

$$P_\Delta\{T(X_1,\cdots,X_m;Y_1,\cdots,Y_n)-\Delta\leqslant t\}$$
$$=P_0\{T(X_1,\cdots,X_m;Y_1,\cdots,Y_n)\leqslant t\}. \quad (7.2.5)$$

上面两个引理的证明与一样本位置参数的情形基本相同.

定理 2.10 设 $\hat{\Delta}(\cdot;\cdot)$是由满足二样本位置参数问题条件(1)～(3)的检验统计量 $V(\cdot;\cdot)$产生的 H-L 估计量. 又满足如下条件：

(i) 总体分布 $F(x)$关于某一 η 对称；

(ii) $V(x_1,\cdots,x_m;y_1,\cdots,y_n)+V(-x_1,\cdots,-x_m;-y_1,\cdots,-y_n)=2\xi$ 对任一$(x_1,\cdots,x_m;y_1,\cdots,y_n)$成立，其中 ξ 是 $V(\cdot;\cdot)$ 在 H_0 下的分布的对称点；

(iii) $V(x_1,\cdots,x_m;y_1,\cdots,y_n)=V(x_1+K,\cdots,x_m+K;y_1+K,\cdots,y_n+K)$对一切$(x_1,\cdots,x_m;y_1,\cdots,y_n)$及 K 成立，

则有

$$\hat{\Delta}(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$$
$$=-\hat{\Delta}(-X_1,\cdots,-X_m;-Y_1,\cdots,-Y_n),$$

且 $\hat{\Delta}(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$的分布关于 Δ 对称.

证明 由定理条件(ii)，与一样本位置参数情况类似，可证

$$\hat{\Delta}(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$$
$$=-\hat{\Delta}(-X_1,\cdots,-X_m;-Y_1,\cdots,-Y_n).$$

由 $\hat{\Delta}(\cdot;\cdot)$是平移的，有

$$\hat{\Delta}(X_1,\cdots,X_m;Y_1,\cdots,Y_n)-\Delta$$
$$=\hat{\Delta}(X_1,\cdots,X_m;Y_1-\Delta,\cdots,Y_n-\Delta).$$

由 $F(x)$ 关于 η 对称，得

$$\begin{aligned}
&\hat{\Delta}(X_1,\cdots,X_m;Y_1-\Delta,\cdots,Y_n-\Delta)\\
&\quad\stackrel{d}{=}\hat{\Delta}(2\eta-X_1,\cdots,2\eta-X_m;\\
&\qquad\quad 2\eta+\Delta-Y_1,\cdots,2\eta+\Delta-Y_n)\\
&\quad=\hat{\Delta}(-X_1,\cdots,-X_m;\Delta-Y_1,\cdots,\Delta-Y_n)\\
&\quad=\Delta+\hat{\Delta}(-X_1,\cdots,-X_m;-Y_1,\cdots,-Y_n)\\
&\quad=\Delta-\hat{\Delta}(X_1,\cdots,X_m;Y_1,\cdots,Y_n).
\end{aligned}$$

上面第二个等号是根据定理的条件(iii). 故有

$$\begin{aligned}
&\hat{\Delta}(X_1,\cdots,X_m;Y_1,\cdots,Y_n)-\Delta\\
&\qquad\stackrel{d}{=}\Delta-\hat{\Delta}(X_1,\cdots,X_m;Y_1,\cdots,Y_n).
\end{aligned}$$

所以 $\hat{\Delta}(\cdot\,;\,\cdot)$ 的分布关于 Δ 对称. 证毕.

例 3 Wilcoxon 秩和统计量 W 满足定理的(i)，(ii)，当总体分布 $F(x)$ 满足(iii)，即 $F(x)$ 是对称分布时，则由 W 产生的 H-L 估计

$$\begin{aligned}
&\hat{\Delta}(X_1,\cdots,X_m;Y_1,\cdots,Y_n)\\
&\qquad=\{Y_j-X_i,\ i=1,\cdots,m,j=1,\cdots,n\}\text{ 的中位数}
\end{aligned}$$

的分布关于参数 Δ 对称.

定理 2.11 $\hat{\Delta}(\cdot\,;\,\cdot)$ 是由满足二样本位置参数问题条件(1)～(3)的检验统计量 $V(\cdot\,;\,\cdot)$ 产生的参数 Δ 的 H-L 估计量. 若

$$P_0\{V(X_1,\cdots,X_m;Y_1,\cdots,Y_n)=\xi\}=0,$$

则 $\hat{\Delta}(\cdot\,;\,\cdot)$ 是 Δ 的中位数无偏估计. 其中 $P_0\{\quad\}$ 表示在 $\Delta=0$ 时计算概率，ξ 是 $\Delta=0$ 时，$V(\cdot\,;\,\cdot)$ 分布的对称点.

证明 与一样本位置参数问题时的证明步骤类似.

§7.3 Hodges-Lehmann 估计量的渐近性质

设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布为 $F(x-\theta)$，函数 $F(t)$ 连续，且关于 0 点对称. $V(X_1,\cdots,X_n)$ 是满足一样本位置

参数问题条件(1)～(3)的检验统计量. $\hat{\theta}(X_1,\cdots,X_n)$是由 $V(\cdot)$ 产生的 θ 的 H-L 估计.

引理 3.1 对任一常数 a,记 $\theta_n=-\frac{a}{\sqrt{n}}$,其中 n 为样本量,又当样本量为 n 时,在 $\theta=0$ 下,统计量 $V(\cdot)$分布的对称点记为 ξ_n. 若在 $\theta=\theta_n$ 下的概率满足

$$\lim_{n\to\infty}P_{\theta_n}\{V(X_1,\cdots,X_n)\leqslant\xi_n\}\text{ 存在}$$

且

$$\lim_{n\to\infty}P_{\theta_n}\{V(X_1,\cdots,X_n)=\xi_n\}=0,$$

则对任一固定的 θ,有

$$\lim_{n\to\infty}P_{\theta}\{\sqrt{n}(\hat{\theta}(X_1,\cdots,X_n)-\theta)\leqslant a\}$$
$$=\lim_{n\to\infty}P_{\theta_n}\{V(X_1,\cdots,X_n)\leqslant\xi_n\}. \qquad (7.3.1)$$

证明 由 $\hat{\theta}(\cdot)$是平移的,知

$$\lim_{n\to\infty}P_{\theta}\{\sqrt{n}(\hat{\theta}(X_1,\cdots,X_n)-\theta)\leqslant a\}$$
$$=\lim_{n\to\infty}P_{\theta}\left\{\hat{\theta}(X_1,\cdots,X_n)-\theta\leqslant\frac{a}{\sqrt{n}}\right\}$$
$$=\lim_{n\to\infty}P_{0}\left\{\hat{\theta}(X_1,\cdots,X_n)\leqslant\frac{a}{\sqrt{n}}\right\}.$$

又由第 2 节引理 2.5,有

$$P_0\left\{V\left(X_1-\frac{a}{\sqrt{n}},\cdots,X_n-\frac{a}{\sqrt{n}}\right)<\xi_n\right\}$$
$$\leqslant P_0\left\{\hat{\theta}(X_1,\cdots,X_n)\leqslant\frac{a}{\sqrt{n}}\right\}$$
$$\leqslant P_0\left\{V\left(X_1-\frac{a}{\sqrt{n}},\cdots,X_n-\frac{a}{\sqrt{n}}\right)\leqslant\xi_n\right\}.$$

而当 $X_i\sim F(x)$时,$X_i-\frac{a}{\sqrt{n}}\sim F\left(x-\left(-\frac{a}{\sqrt{n}}\right)\right)$,故有

$$\lim_{n\to\infty}P_0\left\{V\left(X_1-\frac{a}{\sqrt{n}},\cdots,X_n-\frac{a}{\sqrt{n}}\right)=\xi_n\right\}$$

$$= \lim_{n\to\infty} P_{\theta_n}\{V(X_1,\cdots,X_n) = \xi_n\} = 0.$$

所以

$$\lim_{n\to\infty} P_0\left\{\hat{\theta}(X_1,\cdots,X_n) \leqslant \frac{a}{\sqrt{n}}\right\}$$

$$= \lim_{n\to\infty} P_0\left\{V\left(X_1 - \frac{a}{\sqrt{n}},\cdots,X_n - \frac{a}{\sqrt{n}}\right) \leqslant \xi_n\right\}$$

$$= \lim_{n\to\infty} P_{\theta_n}\{V(X_1,\cdots,X_n) \leqslant \xi_n\},$$

从而引理得证. 证毕.

例 1 考虑 Wilcoxon 符号秩统计量 $W^+(\cdot)$.

W^+ 在 $\theta=0$ 下,分布关于 $\xi_n = \frac{n(n+1)}{4}$ 对称,由它产生的 θ 的 H-L 估计为

$$\hat{\theta}(X_1,\cdots,X_n) = \left\{\frac{X_i + X_j}{2}, i \leqslant j, i, j = 1,\cdots,n\right\}\text{的中位数.}$$

对任一 θ,

$$\left[\frac{W^+(\cdot)}{\binom{n}{2}} - \mu_n(\theta)\right]\Big/\sigma_n(\theta)\ \text{有极限分布}\ N(0,1),$$

其中

$$\mu_n(\theta) = P_\theta\{X_1 + X_2 > 0\} = \int_{-\infty}^{+\infty}[1 - F(-y - 2\theta)]\mathrm{d}F(y),$$

$$\sigma_n(\theta) = \frac{1}{\sqrt{3n}},\quad \mu_n(0) = \frac{1}{2},$$

$$\xi_n = \frac{n(n+1)}{4} = \frac{1}{2}\,\frac{n(n+1)}{2} = \binom{n}{2}\mu_n(0) + \frac{n}{2}.$$

当 $\theta_n = -\frac{a}{\sqrt{n}}$ 时,有

$$\frac{\sqrt{n}}{\binom{n}{2}}\left[\binom{n}{2}\mu_n(\theta_n) - \xi_n\right] = \frac{\mu_n(\theta_n) - \mu_n(0)}{1/\sqrt{n}} + \frac{n^{3/2}}{n(n-1)}.$$

所以

$$\lim_{n\to\infty}\frac{\sqrt{n}}{\binom{n}{2}}\left[\binom{n}{2}\mu_n(\theta_n)-\xi_n\right]=-a\lim_{n\to\infty}\left[\frac{\mu_n\left(\frac{-a}{\sqrt{n}}\right)-\mu_n(0)}{\left(\frac{-a}{\sqrt{n}}\right)}\right]$$

$$=-a\lim_{n\to\infty}\mu_n'(0)=-a\lim_{n\to\infty}2\int_{-\infty}^{+\infty}f^2(y)\mathrm{d}y$$

$$=-2a\int_{-\infty}^{+\infty}f^2(y)\mathrm{d}y.$$

而

$$\frac{\sqrt{n}}{\binom{n}{2}}\left[W^+-\binom{n}{2}\mu_n(\theta_n)\right]\text{ 有极限分布 }N\left(0,\frac{1}{3}\right),$$

所以

$$\frac{\sqrt{n}}{\binom{n}{2}}[W^+-\xi_n]\text{ 有极限分布 }N\left(-2a\int_{-\infty}^{+\infty}f^2(y)\mathrm{d}y,\frac{1}{3}\right).$$

由此得

$$P_{\theta_n}\{W^+\leqslant\xi_n\}=P_{\theta_n}\left\{\frac{\sqrt{n}}{\binom{n}{2}}(W^+-\xi_n)\leqslant 0\right\}$$

$$\to\int_{-\infty}^{0}\frac{\sqrt{3}}{\sqrt{2\pi}}\exp\left[-\frac{3}{2}\left(x+2a\int_{-\infty}^{+\infty}f^2(y)\mathrm{d}y\right)^2\right]\mathrm{d}x$$

$$=\int_{-\infty}^{\frac{aB}{A}}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}t^2}\mathrm{d}t,$$

其中 $A=\sqrt{\frac{1}{3}}$, $B=2\int_{-\infty}^{+\infty}f^2(y)\mathrm{d}y$. 从而由引理 3.1 得

$$\sqrt{n}(\hat{\theta}(X_1,\cdots,X_n)-\theta_n)\text{ 有极限分布 }N\left(0,\left(\frac{A}{B}\right)^2\right).$$

定理 3.2 设 $\hat{\theta}(X_1,\cdots,X_n)$是由满足一样本位置参数问题条

件(1)～(3)的检验统计量 $V(X_1,\cdots,X_n)$ 产生的 θ 的 H-L 估计，若 $V(\cdot)$ 对 $\{\mu_n(\theta)\}$ 及 $\{\sigma_n^2(\theta)\}$ 满足第五章第 1 节定理 1.1(Noether 定理)的条件(A_1)～(A_6)，且

$$\lim_{n\to\infty}\frac{\mu_n(0)-\xi_n}{\sigma_n(0)}=0, \tag{7.3.2}$$

其中 ξ_n 是检验统计量 $V(X_1,\cdots,X_n)$ 在 $\theta=0$ 下的分布的对称点，则 $\sqrt{n}(\hat{\theta}(X_1,\cdots,X_n)-\theta)$ 的极限分布是期望为 0、方差为 $\frac{1}{K_V^2}$ 的 $H(\cdot)$ 分布，其中 K_V 是 $V(\cdot)$ 的检验效力因子，即

$$\lim_{n\to\infty}\frac{\mu_n'(0)}{\sqrt{n\sigma_n^2(0)}}=K_V,$$

$H(\cdot)$ 为 Noether 定理中之极限分布.

证明 对任一 a，令 $\theta_n=-\frac{a}{\sqrt{n}}$，则

$$\begin{aligned}\frac{V-\xi_n}{\sigma_n(\theta_n)}&=\frac{V-\mu_n(\theta_n)}{\sigma_n(\theta_n)}+\frac{\mu_n(\theta_n)-\mu_n(0)}{\sigma_n(\theta_n)}+\frac{\mu_n(0)-\xi_n}{\sigma_n(\theta_n)}\\&=\frac{V-\mu_n(\theta_n)}{\sigma_n(\theta_n)}+\frac{\mu_n'(\theta^*)}{\sigma_n(\theta_n)}\left(-\frac{a}{\sqrt{n}}\right)+\frac{\mu_n(0)-\xi_n}{\sigma_n(\theta_n)},\end{aligned}$$

其中 $0<\theta^*<\theta_n$. 故由 Noether 定理的条件及

$$\lim_{n\to\infty}\frac{\mu_n(0)-\xi_n}{\sigma_n(0)}=0$$

知

$$\frac{V-\xi_n}{\sigma_n(\theta_n)}\ \text{有极限分布}\ H(-aK_V,1).$$

由于 $H(\cdot)$ 连续，故引理 3.1 的条件满足，从而

$$\begin{aligned}\lim_{n\to\infty}P_\theta\{\sqrt{n}(\hat{\theta}-\theta)\leqslant a\}&=\lim_{n\to\infty}P_{\theta_n}\{V-\xi_n\leqslant 0\}\\&=P\{H\leqslant 0\}=P\{H-(-aK_V)\leqslant aK_V\}\\&=P\left\{\frac{H-(-aK_V)}{K_V}\leqslant a\right\},\end{aligned}$$

其中 H 为有分布 $H(-aK_V,1)$的随机变量. 故 $\sqrt{n}(\hat{\theta}-\theta)$的极限分布为 $H\left(0,\frac{1}{K_V^2}\right)$. 证毕.

定义 3.1 设 $\hat{\eta}_1,\hat{\eta}_2$ 是参数 η 的两个渐近无偏估计,即 $\sqrt{n}(\hat{\eta}_1-\eta)$, $\sqrt{n}(\hat{\eta}_2-\eta)$有均值为 0 的渐近分布,且

$$\lim_{n\to\infty}\mathrm{var}(\sqrt{n}\,\hat{\eta}_1)=\sigma_1^2,\quad \lim_{n\to\infty}\mathrm{var}(\sqrt{n}\,\hat{\eta}_2)=\sigma_2^2,$$

则

$$\mathrm{ARE}(\hat{\eta}_1,\hat{\eta}_2)=\frac{\sigma_2^2}{\sigma_1^2} \tag{7.3.3}$$

称为 $\hat{\eta}_1$ 相对于 $\hat{\eta}_2$ 的**渐近相对效率**.

定理 3.3 设 $\hat{\theta}_1(X_1,\cdots,X_n)$和 $\hat{\theta}_2(X_1,\cdots,X_n)$分别是基于 $V_1(X_1,\cdots,X_n)$和 $V_2(X_1,\cdots,X_n)$统计量的 θ 的 H-L 估计. $V_1(\cdot)$, $V_2(\cdot)$均满足定理 3.2 的条件,则渐近相对效率

$$\mathrm{ARE}(\hat{\theta}_1,\hat{\theta}_2)=\mathrm{ARE}(V_1,V_2)=\left(\frac{K_{V_1}}{K_{V_2}}\right)^2.$$

证明 显然.

例 2 考虑下列三个 H-L 估计:

$T=\dfrac{\overline{X}-\theta}{S/\sqrt{n}}$ 产生 H-L 估计: $\hat{\theta}_1=\overline{X}$;

$W^+=\sum\limits_{i=1}^{n}\Psi_iR_i^+$ 产生 H-L 估计:

$$\hat{\theta}_2=\left\{\frac{X_i+X_j}{2},\ i\leqslant j,i,j=1,\cdots,n\right\}\text{的中位数};$$

$B=\sum\limits_{i=1}^{n}\Psi_i$产生 H-L 估计: $\hat{\theta}_3=\{X_i,\ i=1,\cdots,n\}$的中位数,

则

$$\mathrm{ARE}(\hat{\theta}_2,\hat{\theta}_1)=\mathrm{ARE}(W^+,T)=12\sigma^2\left(\int_{-\infty}^{+\infty}f^2(x)\mathrm{d}x\right)^2;$$

$$\mathrm{ARE}(\hat{\theta}_3,\hat{\theta}_1)=\mathrm{ARE}(B,T)=4\sigma^2(f(0))^2;$$

$$\mathrm{ARE}(\hat{\theta}_3,\hat{\theta}_2)=\mathrm{ARE}(B,W^+)=\left(f(0)\Big/\sqrt{3}\int_{-\infty}^{+\infty}f^2(x)\mathrm{d}x\right)^2.$$

对于二样本位置参数问题有类似的定理.

引理 3.4 对任一常数 a，记 $\Delta_N=-\frac{a}{\sqrt{N}}$，$N=m+n$ 为样本总量，$\frac{m}{N}\to\lambda$，$0<\lambda<1$. 又统计量 $V(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$ 满足二样本位置参数问题的条件(1)～(3)，在 $\Delta=0$ 时，$V(\cdot;\cdot)$的分布的对称点为 ξ_N. 若

$$\lim_{N\to\infty}P_{\Delta_N}\{V(X_1,\cdots,X_m;Y_1,\cdots,Y_n)\leqslant\xi_N\}\ 存在；$$

$$\lim_{N\to\infty}P_{\Delta_N}\{V(X_1,\cdots,X_m;Y_1,\cdots,Y_n)=\xi_N\}=0,$$

则对任一固定的 Δ，有

$$\lim_{N\to\infty}P_\Delta\{\sqrt{n}[\hat{\Delta}(X_1,\cdots,X_m;Y_1,\cdots,Y_n)-\Delta]\leqslant a\}$$
$$=\lim_{N\to\infty}P_{\Delta_N}\{V(X_1,\cdots,X_m;Y_1,\cdots,Y_n)\leqslant\xi_N\},\quad(7.3.4)$$

其中 $\hat{\Delta}(\cdot;\cdot)$是基于 $V(\cdot;\cdot)$的 Δ 的 H-L 估计.

定理 3.5 设 $\hat{\Delta}(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$是基于满足二样本位置问题条件(1)～(3)的检验统计量 $V(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$的 Δ 的 H-L 估计，若统计量 $V(\cdot;\cdot)$对$\{\mu_N(\Delta)\}$，$\{\sigma_N^2(\Delta)\}$满足第五章第 1 节 Noether 定理的条件(A_1)～(A_6)，且

$$\lim_{N\to\infty}\frac{\mu_N(0)-\xi_N}{\sigma_N(0)}=0,\qquad(7.3.5)$$

其中 ξ_N 为 $V(\cdot;\cdot)$在 $\Delta=0$ 下的分布的对称点，则 $\sqrt{N}[\hat{\Delta}(X_1,\cdots,X_m;Y_1,\cdots,Y_n)-\Delta]$的渐近分布是期望为 0、方差为$\frac{1}{K_V^2}$的 $H(\cdot)$分布.

定理 3.6 设 $\hat{\Delta}_1(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$和 $\hat{\Delta}_2(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$分别是基于统计量 $V_1(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$和 $V_2(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$的 Δ 的 H-L 估计，统计量 $V_1(\cdot;\cdot)$，$V_2(\cdot;\cdot)$均满足定理 3.5 的条件，则渐近相对效率

$$\mathrm{ARE}(\hat{\Delta}_1,\hat{\Delta}_2)=\mathrm{ARE}(V_1,V_2)=\left(\frac{K_{V_1}}{K_{V_2}}\right)^2.$$

习　题

1. 证明基于符号统计量 $B(X_1,\cdots,X_n)=\sum_{i=1}^{n}\psi(X_i)$的 θ 的 H-L 估计为

$$\hat{\theta}(X_1,\cdots,X_n)=\text{median}(X_1,\cdots,X_n).$$

2. 设 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$是分布 $F(x-\theta)$的样本次序统计量，$F(x-\theta)$连续关于 θ 对称，集合 A 是集合

$$\{(i,j)\,|\,1\leqslant i\leqslant j\leqslant n\}$$

的子集，满足：若 $(i,j)\in A$，则$(n+1-i,n+1-j)\in A$. 令

$$V(X_1,\cdots,X_n)=\sum_{(i,j)\in A}\psi\left(\frac{X_{(i)}+X_{(j)}}{2}\right).$$

求基于$V(X_1,\cdots,X_n)$的 θ 的 H-L 估计. 说明统计量 B 及 W^+均是这类统计量，指出它们对应的 A 集合.

3. 设随机变量 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 分别是分布 $F(x)$和 $F(x-\Delta)$的独立随机样本，$F(x)$连续并关于某一 η 对称，令

$$\overline{Y}=\text{median}(Y_1,\cdots,Y_n),$$

$$V(X_1,\cdots,X_m;Y_1,\cdots,Y_n)=\sum_{i=1}^{m}\psi(\overline{Y}-X_i),$$

求基于$V(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$的 Δ 的 H-L 估计.

4. 设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，分布函数 $F(x-\theta)$连续且关于 θ 对称. 考虑一样本位置参数假设检验问题 H_0: $\theta=0$，H_1: $\theta>0$ 的线性符号秩统计量 $S^+=\sum_{i=1}^{n}\Psi_i a(R_i^+)$. 若 $\hat{\theta}(X_1,\cdots,X_n)$是基于 S^+的 θ 的 H-L 估计，证明 $\hat{\theta}$ 的分布关于 θ 对称.

5. 证明第 2 节引理 2.8.

6. 证明第 2 节引理 2.9.

7. 证明第 2 节定理 2.11.

8. 说明符号统计量 $B=\sum_{i=1}^{n}\psi(X_i)$满足定理 3.2 的条件，求出相应的 H-L 估计的渐近分布.

第八章　影响曲线与稳健估计

在估计总体分布的特征数时，估计量的性质与原总体分布有关. 一般说来，非参数估计对总体分布的依赖比较弱，而一些参数估计则与分布的具体形状有很大关系. 例如样本平均数$\overline{X}$作为正态总体位置参数 μ 的估计，有很好的优良性，但当总体与正态有些差别，特别是在离中心较远处取值的概率有些增大时，亦即所谓"重尾分布"时，其性能会变差. 这样的估计对总体分布的一些"扰动"缺乏抵抗力. 人们在实际应用中当然希望找到能抵抗"扰动"的估计量，希望总体分布与假定的分布有一些差异时，估计量的优良性无大的变化. 这样的估计称之为**稳健估计**. 在讨论稳健性时，影响曲线是一个重要概念.

§8.1　影响曲线

总体分布 $F(x)$的特征数 γ 可以看作是分布 $F(x)$的泛函. 例如，期望值

$$\mu = T_1(F) = \int_{-\infty}^{+\infty} t\mathrm{d}F(t), \tag{8.1.1}$$

中位数　　　$\gamma = T_2(F) = \inf\left\{x: F(x) \geqslant \frac{1}{2}\right\}.$

若以样本经验分布函数 $F_n(x)$替代 $F(x)$，则构成一个统计量，如

$$T_1(F_n) = \int_{-\infty}^{+\infty} t\mathrm{d}F_n(t) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$$

$$T_2(F_n) = \inf\left\{x: F_n(x) \geqslant \frac{1}{2}\right\} = \mathrm{median}(x_1, \cdots, x_n).$$

绝大多数情况在适当的正则条件下，这些统计量服从渐近正态分

布，即 $\sqrt{n}(T(F_n)-T(F))$ 有渐近正态分布 $N(0,V(F))$，这里 $V(F)$ 为 $T(F_n)$ 的渐近方差，$V(F)$ 越小当然越好.

讨论一个统计量 $T(\cdot)$ 对一类分布 $F(x)$ 是否稳健，常用的一个重要概念是影响曲线.

定义 8.1 设 $T(\cdot)$ 是定义在一个分布函数 $F(x)$ 的集合上的泛函，x 是任意实数，$\delta_x(t)$ 是概率为 1 取值 x 的单点分布函数，$T(\cdot)$ 的**影响曲线**定义为

$$\mathrm{IC}_{T,F}(x)=\lim_{\epsilon\to 0+0}\left[\frac{T((1-\epsilon)F(\cdot)+\epsilon\delta_x(\cdot))-T(F(\cdot))}{\epsilon}\right].$$

$\mathrm{IC}_{T,F}(x)$ 曲线反映了总体分布 $F(x)$ 在 x 点分布概率有一点增加时，$T(\cdot)$ 的变化.

对均值统计量 $T_1(\cdot)$，有

$$\mathrm{IC}_{T_1,F}(x)=x-\mu, \tag{8.1.2}$$

其中 $\mu=\int_{-\infty}^{+\infty}t\mathrm{d}F(t)$ 为 $F(x)$ 的期望值. 对中位数统计量 $T_2(\cdot)$，有

$$T((1-\epsilon)F(\cdot)+\epsilon\delta_x(\cdot))$$
$$=\begin{cases}F^{-1}\left(\dfrac{1}{2}-\dfrac{\epsilon}{2(1-\epsilon)}\right), & x<F^{-1}\left(\dfrac{1}{2}\right),\\ 0, & x=F^{-1}\left(\dfrac{1}{2}\right),\\ F^{-1}\left(\dfrac{1}{2}+\dfrac{\epsilon}{2(1-\epsilon)}\right), & x>F^{-1}\left(\dfrac{1}{2}\right),\end{cases}$$

所以

$$\mathrm{IC}_{T_2,F}(x)=\begin{cases}-\dfrac{1}{2f\left(F^{-1}\left(\dfrac{1}{2}\right)\right)}, & x<F^{-1}\left(\dfrac{1}{2}\right),\\ 0, & x=F^{-1}\left(\dfrac{1}{2}\right),\\ \dfrac{1}{2f\left(F^{-1}\left(\dfrac{1}{2}\right)\right)}, & x>F^{-1}\left(\dfrac{1}{2}\right),\end{cases}$$

其中 $f(t)=F'(t)$为密度函数.

可以看出当 x 取远离中心值的极端值,其概率有一点增加时,样本均值比样本中位数的变化要大.这说明样本中位数比样本均值对极端的样本观测值有更好的抗御能力,具有较好的稳健性.但是对于正态分布总体,中位数远不如样本均值有效.人们总是希望寻求稳健性较好,同时对最常遇到的正态等特殊形式的分布又效率较高的估计量.

A. A. Filippova 在论文(Mises theorem on the asymptotic behavior of functionals of empirical distribution functions and its statistical applications, *Theory of probability and its applications*, vol 7(1962), 24～57)中,用泰勒展开证明了,在一般情况下有

$$\sqrt{n}\left[T(F_n)-T(F)-\int_{-\infty}^{+\infty}\mathrm{IC}_{T,F}(t)\mathrm{d}F_n(t)\right]\xrightarrow{P}0,$$

$T(F_n)$的渐近方差

$$V(F)=\int_{-\infty}^{+\infty}[\mathrm{IC}_{T,F}(t)]^2\mathrm{d}F(t). \tag{8.1.3}$$

§8.2 次序统计量的线性组合估计

许多常用的稳健估计是样本次序统计量的线性组合,以 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$记样本 $X_1,\cdots,X_n$ 的次序统计量,

$$T=a_1X_{(1)}+\cdots+a_nX_{(n)}, \tag{8.2.1}$$

其中$\sum_{i=1}^{n}a_i=1$.当取 $a_i=\frac{1}{n}$ $(i=1,\cdots,n)$时,即为样本均值.在上一节我们已经看到对样本的极端观测值,均值是非常敏感的.为使其不太敏感,一个非常简单的方法是对样本观测值进行"修剪",删去那些最极端的观测值.通常规定一个删去的比例 α,因而一个**修剪均值**(或称**截尾均值**)(Trimmed Mean)可表成

$$T(X_1,\cdots,X_n) = \frac{bX_{([\alpha n])+1} + X_{([\alpha n])+2} + \cdots + X_{(n-[\alpha n]-1)} + bX_{(n-[\alpha n])}}{n(1-2\alpha)},$$

其中$[\alpha n]$表示 αn 的整数部分，$b=1+[\alpha n]-\alpha n$ 是一个加权系数．也就是说，如果 αn 是整数，则修剪均值是对样本观测值依大小从两端各删去 αn 个观测后的平均值．若 αn 不是整数，则从两端各删去$[\alpha n]$个观测外，对剩余的观测值中最大值和最小值再给一个删去权数 $\alpha n-[\alpha n]$进行加权平均．特别 $\alpha=0$ 时，修剪均值就是通常的样本均值，$\alpha=0.5$ 时，修剪均值对应的就是样本中位数，$\alpha=0.25$对应的修剪均值一般称为中段均值(midmean)．常用的 α 取值为 0.05 或 0.10.

用泛函的形式表示，修剪均值可写成

$$T_\alpha(F) = \frac{1}{1-2\alpha}\int_\alpha^{1-\alpha} F^{-1}(t)\mathrm{d}t = \frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} t\mathrm{d}F(t),$$

对应的估计量即为

$$T_\alpha(F_n)=\frac{1}{n(1-2\alpha)} \cdot \{bX_{([\alpha n]+1)}+X_{([\alpha n]+2)}+\cdots+X_{(n-[\alpha n]-1)}+bX_{(n-[\alpha n])}\}.$$

α-修剪均值的影响曲线为

$$\mathrm{IC}(x)=\begin{cases}\frac{1}{1-2\alpha}(F^{-1}(\alpha)-C(\alpha)), & x<F^{-1}(\alpha),\\ \frac{1}{1-2\alpha}(x-C(\alpha)), & F^{-1}(\alpha)\leqslant x\leqslant F^{-1}(1-\alpha),\\ \frac{1}{1-2\alpha}(F^{-1}(1-\alpha)-C(\alpha)), & x>F^{-1}(1-\alpha),\end{cases}$$

其中 $$C(\alpha)=\int_\alpha^{1-\alpha} F^{-1}(t)\mathrm{d}t+\alpha(F^{-1}(\alpha)+F^{-1}(1-\alpha)).$$

当 $F(\cdot)$为对称分布时，$C(\alpha)=0$. 这一估计量的渐近方差为

$$V(F) =\frac{1}{(1-2\alpha)^2}\left\{\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} (t-C(\alpha))^2\mathrm{d}F(t) + \alpha[(F^{-1}(\alpha)-C(\alpha))^2+(F^{-1}(1-\alpha)-C(\alpha))^2]\right\}.$$

另一类样本次序统计量线性组合形式的稳健估计为适当选择的线性组合. 如 Gastwirth 在 1966 年提出的,选用中位数和三分之一及三分之二样本分位数的三点平均估计:

$$T_3(X_1,\cdots,X_n)$$
$$= 0.3X_{([\frac{n}{3}+1])} + 0.4\text{median}(X_1,\cdots,X_n) + 0.3X_{(n-[\frac{n}{3}])}$$

以及

$$T_4(X_1,\cdots,X_n)$$
$$= \left[X_{([\frac{n}{4}]+1)} + 2\text{median}(X_1,\cdots,X_n) + X_{(n-[\frac{n}{4}])}\right]\Big/4$$

等等.

这一类估计量写成泛函形式为

$$T(F) = \sum_{i=1}^{m}\beta_i F^{-1}(t_i), \tag{8.2.2}$$

其中 $\sum_{i=1}^{m}\beta_i=1, 0<t_1<\cdots<t_m<1$. 这类估计量的影响曲线在点 $x=F^{-1}(t_i)$ $(i=1,\cdots,m)$ 处有跳跃,跳跃值为 $\beta_i/f(F^{-1}(t_i))$,其他处为常数. 即

$$\text{IC}(x) = \sum \frac{\beta_i}{f(F^{-1}(t_i))} - C(F), \tag{8.2.3}$$

上式中求和号 $\sum$ 是对满足 $\{i: F^{-1}(t_i)<x\}$ 条件的 i 值求和,$C(F)$ 为常数,使 $\text{IC}(x)$ 满足 $\int_{-\infty}^{+\infty}\text{IC}(x)\text{d}F(x)=0$. 这类估计的渐近方差可用下式求出:

$$V(F) = \int_{-\infty}^{+\infty}[\text{IC}(x)]^2\text{d}F(x). \tag{8.2.4}$$

特别,上面提到的 T_3 和 T_4 其 m 均为 3,对 T_3,$(t_1,t_2,t_3)=(1/3,1/2,2/3)$,$(\beta_1,\beta_2,\beta_3)=(0.3,0.4,0.3)$,对 T_4,$(t_1,t_2,t_3)=(1/4,1/2,3/4)$,$(\beta_1,\beta_2,\beta_3)=(1/4,1/2,1/4)$.

当 $F(\cdot)$ 是对称分布时,T_3 和 T_4 的影响曲线为

$$\mathrm{IC}(x)=\begin{cases}-\left(\dfrac{\beta_2}{2f(0)}+\dfrac{\beta_1}{f(F^{-1}(t_1))}\right), & x<F^{-1}(t_1),\\ -\dfrac{\beta_2}{2f(0)}, & F^{-1}(t_1)\leqslant x<0,\\ \dfrac{\beta_2}{2f(0)}, & 0\leqslant x<F^{-1}(t_3),\\ \dfrac{\beta_2}{2f(0)}+\dfrac{\beta_3}{f(F^{-1}(t_3))}, & x\geqslant F^{-1}(t_3).\end{cases}$$

此时，$T_3(X_1,\cdots,X_n)$的渐近方差为

$$V_3(F)=\frac{1}{3}\left(\frac{0.2}{f(0)}\right)^2+\frac{2}{3}\left[\frac{0.2}{f(0)}+\frac{0.3}{f\left(F^{-1}\left(\frac{2}{3}\right)\right)}\right]^2.$$

$T_4(X_1,\cdots,X_n)$的渐近方差为

$$V_4(F)=\frac{1}{32}\left\{\left(\frac{1}{f(0)}\right)^2+\left[\frac{1}{f(0)}+\frac{1}{f\left(F^{-1}\left(\frac{3}{4}\right)\right)}\right]^2\right\}.$$

§8.3 M 估 计

设随机变量 $X_1,\cdots,X_n$ 相互独立同分布，是来自总体 $F(x-\theta)$的样本，$F(x-\theta)$关于 θ 对称. 位置参数 θ 的传统的最小二乘估计方法，是选择 $\theta=\hat{\theta}$，使

$$Q=\sum_{i=1}^{n}(X_i-\theta)^2$$

达到最小值. 可通过正规方程

$$\frac{\partial Q}{\partial\theta}=-2\sum_{i=1}^{n}(X_i-\theta)=0$$

解出 $\hat{\theta}$. 这里采用了平方$(X_i-\theta)^2$ 作为 X_i 与 θ 的距离. 这种获得估计的方法很易推广. 对一般的距离 $\rho(t)$，$\rho(t)$满足距离函数的性质，(i) $\rho(t)\geqslant 0$，对一切 t 成立，且 $\rho(0)=0$；(ii) $\rho(t)=\rho(-t)$，对一切 t. 我们可以选择 $\theta=\hat{\theta}$，使

$$Q = \sum_{i=1}^{n} \rho(X_i - \theta)$$

达到最小值. 当 $\rho(t)$对全部 t 有有限导数 $\rho'(t)$,则 $\hat{\theta}$ 应满足方程

$$\sum_{i=1}^{n} \rho'(X_i - \theta) = 0.$$

也有一些重要的距离 $\rho(t)$,并非对一切 t 都有导数. 如 $\rho(t) = |t|$,在 $t=0$,不存在导数. 要选 $\theta=\hat{\theta}$,使

$$Q = \sum_{i=1}^{n} |X_i - \theta|$$

达到最小值,可如下考虑:设 $X_{(1)} < \cdots < X_{(n)}$是 $X_1, \cdots, X_n$ 的次序统计量,当 $n=2k+1$ 为奇数时,有

$$|X_{(i)} - X_{(k+1)}| + |X_{(k+1)} - X_{(n+1-i)}| = |X_{(i)} - X_{(n+1-i)}|$$

$$\leqslant |X_{(i)} - \theta| + |\theta - X_{(n+1-i)}|, \quad i = 1, \cdots, n.$$

另外亦有 $\quad |X_{(k+1)} - X_{(k+1)}| \leqslant |\theta - X_{(k+1)}|.$

将上列各式相加,即得:对任意 θ 值有

$$\sum_{i=1}^{n} |X_{(i)} - X_{(k+1)}| \leqslant \sum_{i=1}^{n} |X_{(i)} - \theta|,$$

也就是

$$\sum_{i=1}^{n} |X_i - X_{(k+1)}| \leqslant \sum_{i=1}^{n} |X_i - \theta|.$$

所以 $\hat{\theta} = X_{(k+1)}$. 当 $n=2k$ 为偶数时,取 $\hat{\theta} = \frac{1}{2}(X_{(k)} + X_{(k+1)})$,亦会使 Q 达到最小值.

在这种情形,$\hat{\theta}$ 可以看成是满足下列方程的一个解

$$\sum_{i=1}^{n} \psi(X_i - \theta) = 0,$$

其中函数

$$\psi(t) = \begin{cases} -1, & t < 0, \\ 0, & t = 0, \\ 1, & t > 0. \end{cases}$$

但这一方程的解不一定唯一. 解决这一问题的方法是可采用与 Hodges-Lehmann 估计类似的定义.

定义 3.1 设 $\psi(\cdot)$是满足下列性质的函数：

(i) $\psi(t)$是在全直线上定义的一个非降、非常数的函数.

(ii) 对一切 t，有 $\psi(t)=-\psi(-t)$.

由 $\psi(\cdot)$产生的下列估计称为 ***M* 估计**：

$$\hat{\theta}(X_1,\cdots,X_n)=\frac{\theta^*+\theta^{**}}{2}, \tag{8.3.1}$$

其中

$$\theta^*(x_1,\cdots,x_n)=\sup\left\{\theta\left|\sum_{i=1}^n\psi(x_i-\theta)>0\right.\right\},$$

$$\theta^{**}(x_1,\cdots,x_n)=\inf\left\{\theta\left|\sum_{i=1}^n\psi(x_i-\theta)<0\right.\right\}.$$

由于 $\psi(t)$满足性质(i)和(ii)，因而一定存在一个 t_0，使 $t>t_0$ 时，$\psi(t)>0$，而 $t<-t_0$ 时，$\psi(t)<0$. 如此，当 $\theta<X_{(1)}-t_0$ 时，必有

$$\sum_{i=1}^n\psi(X_i-\theta)>0,$$

而当 $\theta>X_{(n)}+t_0$ 时，必有 $\sum_{i=1}^n\psi(X_i-\theta)<0$，从而一定存在有限的 θ^* 和 θ^{**}. 当 $\psi(t)$连续且严格递增时，M 估计 $\hat{\theta}$ 可从下式解出

$$\sum_{i=1}^n\psi(X_i-\theta)=0. \tag{8.3.2}$$

如果总体分布除未知位置参数 θ 外，还有尺度参数 σ，则构造 M 估计，需要 $\psi(\cdot)$和另一个函数 $\eta(\cdot)$，从方程组

$$\begin{cases}\displaystyle\sum_{i=1}^n\psi\left(\frac{X_i-\theta}{\sigma}\right)=0,\\ \displaystyle\sum_{i=1}^n\eta\left(\frac{X_i-\theta}{\sigma}\right)=0\end{cases} \tag{8.3.3}$$

解出 θ 和 σ.

这是一个很广泛的估计类，参数估计中常用的极大似然估计也在其中. 若总体分布有密度函数

$$\frac{1}{\sigma}f\left(\frac{t-\theta}{\sigma}\right),$$

给定样本值 $x_1,\cdots,x_n$,似然函数为

$$L(\theta,\sigma) = \sum_{i=1}^{n}\left[\log f\left(\frac{x_i-\theta}{\sigma}\right) - \log\sigma\right].$$

在适当条件下,可通过正规方程组:

$$\begin{cases}\dfrac{\partial L}{\partial\theta} = \sum_{i=1}^{n}\left[-f'\left(\dfrac{x_i-\theta}{\sigma}\right)\Big/f\left(\dfrac{x_i-\theta}{\sigma}\right)\right]=0,\\[2ex] \dfrac{\partial L}{\partial\sigma} = \sum_{i=1}^{n}\left[-\left(\dfrac{x_i-\theta}{\sigma^2}\right)f'\left(\dfrac{x_i-\theta}{\sigma}\right)\Big/f\left(\dfrac{x_i-\theta}{\sigma}\right)-\dfrac{1}{\sigma}\right]=0\end{cases}$$

解得极大似然估计.当取

$$\psi(t) = -\frac{f'(t)}{f(t)},\quad \eta(t) = -\frac{tf'(t)}{f(t)} - 1$$

时,对应的 M 估计即为极大似然估计.前面我们也已看到最小二乘、最小一乘估计也在 M 估计的范围内,M 估计这一名称正是来自这些事实.

P. J. Huber 在 1964 年提出一类 M 估计.他取

$$\psi(t) = \begin{cases}-k, & t<-k,\\ x, & -k\leqslant t\leqslant k,\\ k, & t>k.\end{cases}\qquad(8.3.4)$$

而取

$$\eta(t) = \psi^2(t) - \beta(k),\qquad(8.3.5)$$

其中 $\beta(k)=\int_{-\infty}^{+\infty}\psi^2(t)\mathrm{d}\Phi(t)$,$\Phi(t)$为标准正态分布函数.解方程组

$$\begin{cases}\sum_{i=1}^{n}\psi\left(\dfrac{x_i-\theta}{\sigma}\right)=0,\\[2ex] \sum_{i=1}^{n}\eta\left(\dfrac{x_i-\theta}{\sigma}\right)=0\end{cases}$$

一般需要通过迭代解出 θ 和 σ,初始值可以取 θ_0 为样本中位数,σ_0 取$(x_{(n)}-x_{(1)})/(1.35)$.$(x_{(n)}-x_{(1)})$为样本差程.

Huber 所取的 $\psi(\cdot)$对应的距离为

$$\rho(t)=\begin{cases}\dfrac{1}{2}t^2, & |t|\leqslant k,\\ k|t|-\dfrac{1}{2}k^2, & |t|>k.\end{cases}\tag{8.3.6}$$

这个函数在中部采用了平方距离，而在两端采用了绝对值距离，因此这一类 M 估计倾向于像取平均一样处理中间大小的样本观测值，而像取中位数一样不重视两边极端的样本观测值.

如果假定总体分布密度为$\frac{1}{\sigma}f\left(\frac{x-\theta}{\sigma}\right)$，位置参数 θ 和尺度参数 σ 的上述这些 M 估计写成泛函 $T(F)$和 $S(F)$，应满足

$$\int_{-\infty}^{+\infty}\psi\left(\frac{t-T(F)}{S(F)}\right)\mathrm{d}F(t)=0,$$

$$\int_{-\infty}^{+\infty}\eta\left(\frac{t-T(F)}{S(F)}\right)\mathrm{d}F(t)=0.$$

将上式中的 $F(x)$替换为 $F_\epsilon(x)=(1-\epsilon)F(x)+\epsilon\delta_x$，然后对 ϵ 求导数，则可发现它们的影响曲线应满足：

$$\begin{aligned}&\mathrm{IC}_{T,F}(x)\,\frac{1}{S(F)}\int_{-\infty}^{+\infty}\psi'\left(\frac{t-T(F)}{S(F)}\right)\mathrm{d}F(t)\\&\qquad+\mathrm{IC}_{S,F}(x)\,\frac{1}{S(F)}\int_{-\infty}^{+\infty}\psi'\left(\frac{t-T(F)}{S(F)}\right)\frac{t-T(F)}{S(F)}\mathrm{d}F(t)\\&=\psi\left(\frac{x-T(F)}{S(F)}\right),\end{aligned}\tag{8.3.7}$$

$$\begin{aligned}&\mathrm{IC}_{T,F}(x)\,\frac{1}{S(F)}\int_{-\infty}^{+\infty}\eta'\left(\frac{t-T(F)}{S(F)}\right)\mathrm{d}F(t)\\&\qquad+\mathrm{IC}_{S,F}(x)\,\frac{1}{S(F)}\int_{-\infty}^{+\infty}\eta'\left(\frac{t-T(F)}{S(F)}\right)\frac{t-T(F)}{S(F)}\mathrm{d}F(t)\\&=\eta\left(\frac{x-T(F)}{S(F)}\right).\end{aligned}\tag{8.3.8}$$

如果 $F(x)$对称，$\psi(\cdot)$为奇函数，$\eta(\cdot)$为偶函数，则简化为

$$\mathrm{IC}_{T,F}(x)=\left[S(F)\Big/\int_{-\infty}^{+\infty}\psi'\left(\frac{t}{S(F)}\right)\mathrm{d}F(t)\right]\psi\left(\frac{x}{S(F)}\right),\tag{8.3.9}$$

$$\mathrm{IC}_{S,F}(x)=\left[S(F)\Big/\int_{-\infty}^{+\infty}\eta'\left(\frac{t}{S(F)}\right)\frac{t}{S(F)}\mathrm{d}F(t)\right]\eta\left(\frac{x}{S(F)}\right). \tag{8.3.10}$$

可以看到，对于对称分布 $F(\cdot)$，上面提到的 Huber 的 M 估计，当取 $\alpha=F(-kS(F))$ 时，其 $T(F)$ 的影响曲线与 § 8.2 节中 α-修剪平均的影响曲线相同.

习　题

1. 当 $n=2k$ 为偶数时，$X_{(1)},\cdots,X_{(n)}$ 为样本 $X_1,\cdots,X_n$ 的次序统计量，m 为区间 $(X_{(k)},X_{(k+1)})$ 内任一数，证明对任意的 θ 值，有

$$\sum_{i=1}^{n}|x_i-m|\leqslant\sum_{i=1}^{n}|x_i-\theta|.$$

2. 具体计算出样本均值和样本中位数的影响曲线.

3. 具体计算 § 8.2 节中 α-修剪平均 $T_\alpha(F)$ 的影响曲线.

4. 设 $\hat{\theta}(X_1,\cdots,X_n)$ 是 § 8.3 中定义 3.1 给出的 M 估计，证明

$$\hat{\theta}(x_1,\cdots,x_n)=-\hat{\theta}(-x_1,\cdots,-x_n).$$

5. 如果总体分布关于位置参数 θ 对称，证明：由 § 8.3 节中定义 3.1 给出的 M 估计 $\hat{\theta}(X_1,\cdots,X_n)$ 有关于 θ 对称的分布.

6. 设 $x_1,\cdots,x_n$ 是来自总体 $F(x)$ 的样本值. 它们构成的经验分布函数记为 $F_n(t)$，现考虑将样本容量增加为 $n+1$，样本值为 $x_1,\cdots,x_n,x$，这 $n+1$ 个样本观测值构成的经验分布函数记为 $F_{n+1}^*(t)$，则对一个泛函 $T(\cdot)$，函数

$$S(x)=T(F_{n+1}^*)-T(F_n)$$

称为 $T(\cdot)$ 的敏感曲线. 它反响新添样本值 x 给估计造成的影响. 试求样本均值、样本中位数的敏感曲线.

7. 若取距离函数 $\rho(t)=t^4$，求对应的位置参数 θ 的 M 估计.

第九章　使用秩统计量的一些其他统计问题

§9.1　k 样本问题(一因素分组问题)

设随机变量 $X_{i1},\cdots,X_{in_i}$相互独立同分布，分布为 $F(x-\theta_i)$，函数 $F(t)$连续，$i=1,\cdots,k,n_1+n_2+\cdots+n_k=N$，检验假设

$$H_0: \theta_1=\theta_2=\cdots=\theta_k;$$

备择假设

$$H_1: 对某些\ i\neq j\ 有\ \theta_i\neq\theta_j,\quad i,j=1,\cdots,k.$$

将全部样本共 N 个随机变量一起排序，由

$$X_{11},\cdots,X_{1n_1};\ X_{21},\cdots,X_{2n_2};\ \cdots;\ X_{k1},\cdots,X_{kn_k}$$

得 k 个独立样本的混合秩统计量

$$R_{11},\cdots,R_{1n_1};\ R_{21},\cdots,R_{2n_2};\ \cdots;\ R_{k1},\cdots,R_{kn_k}.$$

对混合秩统计量有下列定理.

定理 1.1　在 H_0 下(即 k 样本来自同一总体)，以 R_{ij}记 X_{ij}在全部样本混合样本中的秩，又记

$$R_{i\cdot}=\sum_{j=1}^{n_i}R_{ij},\quad \overline{R}_{i\cdot}=\frac{1}{n_i}\sum_{j=1}^{n_i}R_{ij},\tag{9.1.1}$$

则

$$E(R_{i\cdot})=n_i(N+1)/2,\quad E(\overline{R}_{i\cdot})=(N+1)/2,$$

$$\mathrm{var}R_{i\cdot}=n_i(N-n_i)(N+1)/12,$$

$$\mathrm{cov}(R_{i\cdot},R_{j\cdot})=-n_in_j(N+1)/12.$$

证明　直接计算即得.

利用混合秩可构造检验前述假设的检验统计量

$$H = \sum_{i=1}^{k} C_{iN}^2 \left[\frac{\overline{R}_{i.} - E(\overline{R}_{i.})}{\sqrt{\operatorname{var}(\overline{R}_{i.})}} \right]^2. \qquad (9.1.2)$$

显然 H 反映了样本的组间差异，H 取大的值有利于备择假设 H_1. 适当选取 $\{C_{iN}\}$ 可以使 H 在原假设 H_0 下有近似 $k-1$ 自由度的 χ^2 分布.

定理 1.2 设 k 样本来自同一总体，且

$$N \to \infty,\ n_i/N \to \lambda_i,\ 0 < \lambda_i < 1\ (i = 1, \cdots, k).$$

又设 $C_{iN} \to C_i\ (i=1,\cdots,k)$，定义统计量

$$T_i = C_{iN} \frac{1}{\sqrt{N}} (\overline{R}_{i.} - E(\overline{R}_{i.})),\quad i = 1, \cdots, k, \quad (9.1.3)$$

则向量 $T=(T_1,\cdots,T_k)$ 有近似多维正态分布 $\mathrm{MVN}(0,B)$，其中 $B=(b_{ij})$ 是协差阵

$$b_{ij} = \begin{cases} C_i^2(1-\lambda_i)/(12\lambda_i), & \text{当 } i = j \text{ 时}, \\ -C_iC_j/12, & \text{当 } i \neq j \text{ 时}. \end{cases}$$

证明 将随机变量 $X_{ij}\ (i=1,\cdots,k, j=1,\cdots,n_i)$ 写成

$$Y_1, \cdots, Y_N, \quad N = \sum_{i=1}^{k} n_i.$$

以 J 表示第 i 样本 $\{X_{i1},\cdots,X_{in_i}\}$ 在 $\{Y_1,\cdots,Y_N\}$ 中的下标集，$\overline{J}$ 为 J 的余集，则有

$$R_{i.} = \#\{Y_v > Y_u, v \in J, u \in \overline{J}\} + \frac{n_i(n_i+1)}{2}$$

$$= \sum_{v \in J} \sum_{u \in \overline{J}} \psi(Y_v - Y_u) + \frac{n_i(n_i+1)}{2},$$

$$R_{i.} - E(R_{i.}) = \sum_{v \in J} \sum_{u \in \overline{J}} \left[\psi(Y_v - Y_u) - \frac{1}{2} \right].$$

在 H_0 下，有

$$E\left[\psi(Y_v - Y_u) - \frac{1}{2} \middle| Y_i = y \right]$$

$$
= \begin{cases} 0, & u \neq i, v \neq i, \\ F(y) - \dfrac{1}{2}, & v = i, \\ \dfrac{1}{2} - F(y), & u = i. \end{cases}
$$

因此有

$$
\begin{aligned}
& E[R_{i\cdot} - E(R_{i\cdot}) | Y_i = y] \\
& = \sum_{v \in J} \sum_{u \in \bar{J}} E\left[\psi(Y_v - Y_u) - \frac{1}{2} \middle| Y_i = y\right] \\
& = \begin{cases} (N - n_i)\left[F(y) - \dfrac{1}{2}\right], & i \in J, \\ n_i\left[\dfrac{1}{2} - F(y)\right], & i \in \bar{J}. \end{cases}
\end{aligned}
$$

故 $T_i = C_{iN}\dfrac{1}{\sqrt{N}}(\bar{R}_{i\cdot} - E(R_{i\cdot})) = C_{iN}\dfrac{1}{\sqrt{N}}\dfrac{R_{i\cdot} - E(R_{i\cdot})}{n_i}$ 的投影为

$$
\begin{aligned}
V_i &= C_{iN} \frac{1}{\sqrt{N}} \left\{\frac{N - n_i}{n_i} \sum_{i \in J} \left[F(Y_i) - \frac{1}{2}\right] + \sum_{i \in \bar{J}} \left[\frac{1}{2} - F(Y_i)\right]\right\} \\
&= C_{iN} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \left[F(Y_i) - \frac{1}{2}\right],
\end{aligned}
$$

其中

$$
a_i = \begin{cases} (N - n_i)/n_i, & \text{当 } i \in J \text{ 时}, \\ -1, & \text{当 } i \in \bar{J} \text{ 时}. \end{cases}
$$

而

$$
\begin{aligned}
\operatorname{var} V_i &= C_{iN}^2 \frac{1}{N} \frac{1}{12} \left\{ n_i \frac{(N - n_i)^2}{n_i^2} + (N - n_i) \right\} \\
&= C_{iN}^2 \frac{N(N - n_i)}{12 n_i N} \\
&\to C_i^2 (1 - \lambda_i)/(12\lambda_i), \\
\operatorname{cov}(V_i, V_j) &\to -C_i C_j / 12.
\end{aligned}
$$

由中心极限定理知 $V = (V_1, \cdots, V_k)$ 是服从渐近多维正态分布

MVN$(0,B)$. 又由投影的性质知

$$E(T_i - V_i)^2 = \mathrm{var}T_i - \mathrm{var}V_i,$$

而

$$\mathrm{var}T_i = C_{iN}^2 \frac{1}{N} \frac{(N-n_i)(N+1)}{12n_i}$$
$$\to C_i^2(1-\lambda_i)/(12\lambda_i).$$

故知$(T-V)$依概率收敛到 0,从而 T 与 V 有相同的极限分布 MVN$(0,B)$. 证毕.

定理 1.3 设随机向量$(Z_1,\cdots,Z_k)$有 MVN$(0,A)$分布,且其中 A 满足 $A^2=A$, $\mathrm{rank}(A)=r$,则统计量 $\sum_{i=1}^k Z_i^2$ 遵从 $\chi^2(r)$分布.

证明 存在正交阵 G,使 $G'AG$ 为对角阵,对角线上有 r 个 1, $k-r$ 个 0. 作随机变量变换

$$U = G'Z, \quad Z = (Z_1,\cdots,Z_k)',$$

则

$$U \sim \mathrm{MVN}(0,G'AG).$$

而

$$\sum_{i=1}^k Z_i^2 = Z'Z = U'G'GU = U'U = \sum_{i=1}^k U_i^2,$$

因而有分布 $\chi^2(r)$. 证毕.

定理 1.4 在 H_0 下(即 k 样本来自同一总体),则当 $\frac{n_i}{N}\to\lambda_i$, $0<\lambda_i<1$ $(i=1,\cdots,k)$, $N\to\infty$时,

$$H^* = \sum_{i=1}^k \left(1-\frac{n_i}{N}\right)\left[\frac{\overline{R}_{i\cdot} - \frac{N+1}{2}}{\sqrt{(N-n_i)(N+1)/(12n_i)}}\right]^2$$
$$= \frac{12}{N(N+1)}\sum_{i=1}^k n_i\left[\overline{R}_{i\cdot} - \frac{N+1}{2}\right]^2 \qquad (9.1.4)$$

有渐近分布 $\chi^2(k-1)$.

证明 取 $C_{iN}=\left(\frac{12n_i}{N+1}\right)^{\frac{1}{2}}$，则 $C_{iN}\to\sqrt{12\lambda_i}=C_i$. 由定理 1.2 知，取

$$T_i = C_{iN}\frac{1}{\sqrt{N}}\left(\overline{R}_{i\cdot}-\frac{N+1}{2}\right)=\sqrt{\frac{12n_i}{N(N+1)}}\left(\overline{R}_{i\cdot}-\frac{N+1}{2}\right),$$

则 $T=(T_1,\cdots,T_k)$ 有渐近分布 MVN$(0,B)$，此处 $B=(b_{ij})$，

$$b_{ij}=\begin{cases}C_i^2\dfrac{1-\lambda_i}{12\lambda_i}=1-\lambda_i, & \text{当 } i=j \text{ 时},\\ -\dfrac{C_iC_j}{12}=-\sqrt{\lambda_i\lambda_j}, & \text{当 } i\neq j \text{ 时}.\end{cases}$$

易验证 B 满足 $B^2=B$，且 $\mathrm{rank}(B)=k-1$，而

$$H^*=\frac{12}{N(N+1)}\sum_{i=1}^{k}n_i\left[\overline{R}_{i\cdot}-\frac{n+1}{2}\right]^2=\sum_{i=1}^{k}T_i^2.$$

由定理 1.3 即得定理的结论. 证毕.

统计量 H^* 称为 Kruskal-Wallis 统计量，本节开头所述 k 样本问题的检验常用此统计量. 否定域为 $\{H^*>h_\alpha\}$. 临界值 h_α 在小样本时由精确分布计算. 在 H_0 下混合秩统计量

$$R=(R_{11},\cdots,R_{1n_1},\cdots,R_{k1},\cdots,R_{kn_k})$$

在集合 $\mathscr{R}=\{(1,2,\cdots,N)$ 的全部排列$\}$ 上均匀分布. 在较小的 k 及 $n_1,\cdots,n_k$ 时，有 H^* 的临界值表. 当 N 较大(如 $k>3$，$n_i\geqslant 5$，$i=1,\cdots,k$)时，一般利用近似分布 $\chi^2(k-1)$ 确定统计量 H^* 的临界值 h_α.

当假设检验问题为

$$H_0:\ \theta_1=\theta_2=\cdots=\theta_k;$$

备择假设

$H_1:\ \theta_1\leqslant\theta_2\leqslant\cdots\leqslant\theta_k$，且至少有一个严格不等式成立

这种形式时，Jonckheere-Terpstra 建议了一个统计量，它与二样本问题的 Mann-Whitney 统计量类似. 令

$$U_{ij}=\sum_{s=1}^{n_i}\sum_{t=1}^{n_j}\phi(X_{jt}-X_{is}),\quad i<j,\ i,j=1,\cdots,k. \tag{9.1.5}$$

Jonckheere-Terpstra 统计量为

$$J = \sum_{i<j}\sum U_{ij}, \tag{9.1.6}$$

显然 J 取大的值有利 H_1. 当 $N\to\infty, \frac{n_i}{N}\to\lambda_i, 0<\lambda_i<1\ (i=1,\cdots,k)$ 时，在 H_0 下

$$\frac{J - E(J)}{\sqrt{\mathrm{var}(J)}} \text{ 有渐近正态分布 } N(0,1),$$

其中

$$E(J)=\frac{N^2-\sum_{i=1}^{k}n_i^2}{4}, \quad \mathrm{var}(J)=\frac{N^2(2N+3)-\sum_{i=1}^{k}n_i^2(2n_i+3)}{72}.$$

故样本量较小时，可由专用临界值表查得否定域的临界值. 样本量较大时用近似正态分布确定临界值.

§9.2 二因素分组问题

二因素分组问题的数学模型有两种提法.

模型Ⅰ 有 n 个区组，k 个处理的观测

$$\{X_{ij}\}, \quad i = 1,\cdots,n,\ j = 1,\cdots,k,$$

全部 X_{ij}相互独立，同一区组内的处理观测 $X_{i1},\cdots,X_{ik}$有相同的分布类型，但位置参数可能不同，不同的区组分布不相同，即 X_{ij}有分布 $F_i(x-\theta_j)\ (j=1,\cdots,k,\ i=1,\cdots,n)$，分布 $F_i(t)\ (i=1,\cdots,n)$连续.

模型Ⅱ 有 n 个区组，k 个处理的观测

$$\{X_{ij}\}, \quad i = 1,\cdots,n,\ j = 1,\cdots,k,$$

各区组的观测$(X_{i1},\cdots,X_{ik})\ (i=1,\cdots,n)$相互独立，有共同的 k 维分布 $F(x_1-\theta_1,\cdots,x_k-\theta_k)$，分布函数 $F(t_1,\cdots,t_k)$连续是变量可置换的函数，即对于$(t_1,\cdots,t_k)$的任一排列$(t_{S_1},\cdots,t_{S_k})$均有

$$F(t_{S_1},\cdots,t_{S_k}) = F(t_1,\cdots,t_k).$$

对模型Ⅰ或Ⅱ检验假设

$$H_0: \theta_1 = \theta_2 = \cdots = \theta_k,$$

备择假设

$$H_1: \text{对某些 } i \neq j \text{ 有 } \theta_i \neq \theta_j, \quad i,j = 1,\cdots,k.$$

首先对每一 i $(i=1,\cdots,n)$，当 H_0 成立时，随机变量

$$X_{i1},\cdots,X_{ik}$$

排序，得到对应的秩向量 $R_{i1},\cdots,R_{ik}$，则秩向量

$$R_i = (R_{i1},\cdots,R_{ik})$$

在集合 $\mathscr{R}=\{(1,\cdots,k)\text{的全部排列}\}$上均匀分布. 因此由它们构成的统计量是非参数适应任意分布的. 令

$$R_{\cdot j} = \sum_{i=1}^{n} R_{ij}, \quad \overline{R}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} R_{ij}, \tag{9.2.1}$$

有

$$E(R_{\cdot j}) = \frac{n(k+1)}{2}, \quad \operatorname{var}(R_{\cdot j}) = \frac{n(k^2-1)}{12},$$

$$\operatorname{cov}(R_{\cdot i}, R_{\cdot j}) = -\frac{n(k+1)}{12},$$

且有

$$\overline{R} = \frac{1}{k}\sum_{j=1}^{k} \overline{R}_{\cdot j} = \frac{1}{nk}\sum_{j=1}^{k}\sum_{i=1}^{n} R_{ij} = \frac{1}{k}\sum_{j=1}^{k} j = \frac{k+1}{2},$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k}(R_{ij} - \overline{R})^2 = n\sum_{j=1}^{k}\left(j - \frac{k+1}{2}\right)^2 = \frac{nk(k^2-1)}{12}.$$

Friedman 等建议用下列统计量检验前述假设：

$$\begin{aligned} S &= \sum_{j=1}^{k}\left(\frac{k-1}{k}\right)\left[\frac{R_{\cdot j} - E(R_{\cdot j})}{\sqrt{\operatorname{var}(R_{\cdot j})}}\right]^2 \\ &= \sum_{j=1}^{k}\left(\frac{k-1}{k}\right)\left[\frac{R_{\cdot j} - n(k+1)/2}{\sqrt{n(k^2-1)/12}}\right]^2 \\ &= \frac{12}{nk(k+1)}\sum_{j=1}^{k}\left[R_{\cdot j} - \frac{n(k+1)}{2}\right]^2. \end{aligned} \tag{9.2.2}$$

这实际上相当于用秩数据$\{R_{ij}, i=1,\cdots,n, j=1,\cdots,k\}$作方差分

析，实际上有

$$S = \frac{\sum_{j=1}^{k} n(\overline{R}_{\cdot j} - \overline{R})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n}(R_{ij} - \overline{R})^2/n(k-1)}.$$

显然 S 取大的值有利于 H_1，否定域应为

$$S > \zeta_\alpha,$$

其中 ζ_α 是相应于检验水平 α 的临界值. 当 n,k 较小时，可由古典概率直接计数统计量的概率分布，有供实际应用的临界值表. 当 $n \to \infty$ 时，Friedman 统计量 S 有渐近 $\chi^2(k-1)$ 分布. 当 n 较大时，可用它确定近似的临界值.

当检验的假设是

$$H_0: \theta_1 = \theta_2 = \cdots = \theta_k,$$

备择假设

$H_1: \theta_1 \leqslant \theta_2 \leqslant \cdots \leqslant \theta_k$，且至少有一个严格不等式成立

这种形式时，用下述 Page 统计量有更好的功效，

$$V = \frac{1}{\sqrt{n}}\sum_{j=1}^{k}\left(j - \frac{k+1}{2}\right)\left(R_{\cdot j} - \frac{n(k+1)}{2}\right), \quad (9.2.3)$$

这是 n 个线性秩统计量的线性组合. 它与统计量

$$L = \sum_{j=1}^{k} jR_{\cdot j} \quad (9.2.4)$$

是等价的检验统计量. 当 V 或 L 取大的值时有利于 H_1，否定域的形式为 $\{L > l_\alpha\}$. 当 k,n 较小时，可直接用古典概率方法计数确定临界值 l_α，有临界值表可供实际使用. 当 $n \to \infty$ 时，在 H_0 下，显然有

$$\frac{L - E(L)}{\sqrt{\mathrm{var}(L)}} \text{ 有渐近正态分布 } N(0,1),$$

这里 $E(L) = nk(k+1)^2/4$，$\mathrm{var}(L) = nk^2(k^2-1)(k+1)/144$. 在 n 较大时一般按正态分布确定近似的临界值 l_α.

§ 9.3 相关性检验

我们考虑一对随机变量(X,Y)的相关问题. 从一个二维连续分布总体取得独立同分布样本$(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)$. 一个通常会碰到的实际问题是: X 和 Y 是否独立,即检验假设

$$H_0:\ X \text{ 和 } Y \text{ 独立}.$$

备择假设可以是单边的(X 和 Y 有正相关,或 X 和 Y 有负相关),也可以是双边的(X 和 Y 相关). 此处以单边备择假设为例进行讨论,

$$H_1:\ X \text{ 和 } Y \text{ 有正相关}.$$

不妨设样本 $X_1<X_2<\cdots<X_n$,则 $X_1,X_2,\cdots,X_n$ 对应的秩向量为$(1,2,\cdots,n)$,而相应的 $Y_1,Y_2,\cdots,Y_n$ 的秩记为 $R_1,R_2,\cdots,R_n$. 当 X 和 Y 有正相关时,样本值 X_i,Y_i 会存在同时取大值或同时取小值的倾向,因而秩统计量对$(1,R_1),(2,R_2),\cdots,(n,R_n)$的相关系数会取大的值,即

$$S=\frac{\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)\left(R_i-\frac{n+1}{2}\right)}{\sqrt{\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)^2\cdot\sum_{i=1}^{n}\left(R_i-\frac{n+1}{2}\right)^2}}=\frac{\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)R_i}{n(n^2-1)/12}$$

$$=1-\frac{6}{n(n^2-1)}\sum_{i=1}^{n}(i-R_i)^2. \tag{9.3.1}$$

取大的值有利于 H_1. 统计量 S 称为 Spearman 秩相关系数,用它作检验统计量检验上述(H_0,H_1),其否定域形式应为$\{S\geqslant s_\alpha\}$, s_α 为显著性水平为 α 的临界值. 对于双边备择假设 H_1,否定域形式为$\{|S|\geqslant s_\alpha\}$. 从上面统计量 S 的第二个等式可以看出 S 是一个线性秩统计量. 故否定域临界值 s_α 可用线性秩统计量的分布求得. 在 H_0 下有

$$E(S)=0,\quad \mathrm{var}(S)=\frac{1}{n-1}.$$

当 $n\to\infty$时，$\sqrt{n-1}S$ 有渐近分布 $N(0,1)$. 对较小的 n，有专用的 Spearman 秩相关统计量临界值表，可查得 s_α，对较大的 n 则用正态分布求得近似的临界值.

对上述独立性假设(H_0,H_1)，另一个著名的非参数检验方法为 Kendall 检验. 其方法如下：对样本$(X_1,Y_1),\cdots,(X_n,Y_n)$，定义

$$\xi(X_i,X_j;Y_i,Y_j)=\operatorname{sgn}(X_j-X_i)\operatorname{sgn}(Y_j-Y_i),$$

其中

$$\operatorname{sgn}(t)=\begin{cases}1, & t>0,\\ 0, & t=0,\\ -1, & t<0.\end{cases}$$

Kendall 统计量为

$$K=\frac{2}{n(n-1)}\sum_{i<j}^{n}\sum^{n}\xi(X_i,X_j;Y_i,Y_j), \qquad (9.3.2)$$

其取值范围从-1到 1，当 X 样本和 Y 样本的大小顺序完全一致时 $K=1$. 若假定 $X_1<\cdots<X_n$，则有

$$\begin{aligned}K&=1-\frac{4}{n(n-1)}\sum_{i<j}\sum\psi(Y_i-Y_j)\\&=1-\frac{4}{n(n-1)}\sum_{i<j}\sum\psi(R_i-R_j).\end{aligned}$$

由此有

$$\begin{aligned}E(K)&=\frac{2}{n(n-1)}\sum_{i<j}\sum E[\xi(X_i,X_j;Y_i,Y_j)]\\&=E\{\operatorname{sgn}(X_2-X_1)\operatorname{sgn}(Y_2-Y_1)\}\\&=P\{(X_2-X_1)(Y_2-Y_1)>0\}\\&\quad-P\{(X_2-X_1)(Y_2-Y_1)<0\}\\&=1-2P\{(X_2-X_1)(Y_2-Y_1)<0\}.\end{aligned}$$

K 是(2,2)可估参数 $E(K)$的 U 统计量，因而有渐近正态分布. 在 H_0 下，有

$$E(K)=0,\quad \operatorname{var}(K)=\frac{2(2n+5)}{9n(n-1)},$$

$$\frac{K-E(K)}{\sqrt{\operatorname{var}K}}\text{ 有渐近分布 }N(0,1).$$

对正相关的备择假设 H_1，显然 K 取大的值有利于 H_1，故否定域有形式 $\{K \geqslant k_a\}$. k_a 为对应显著水平 a 的临界值，在 n 较小时可在 Kendall 相关性检验统计用表中查得，n 较大时则由正态分布确定近似值.

在假定 $X_1 < \cdots < X_n$ 下，实际上有 Spearman 统计量

$$S = 1 - \frac{12}{n(n^2-1)} \sum_{i<j}\sum (j-i)\psi(R_i - R_j). \quad (9.3.3)$$

而 Kendall 统计量

$$K = 1 - \frac{4}{n(n-1)} \sum_{i<j}\sum \psi(R_i - R_j). \quad (9.3.4)$$

比较两者，可以看到统计量 S 给与间隔远的 $(X_i, X_j; Y_i, Y_j)$ 的不一致以更大的权，而 K 统计量的权与间隔远近无关.

二因素分组问题中的 n 个统计量与秩相关系数有密切联系. Page 统计量

$$\begin{aligned} V &= \frac{1}{\sqrt{n}} \sum_{j=1}^{k} \left(j - \frac{k+1}{2} \right) \left(R_{\cdot j} - \frac{n(k+1)}{2} \right) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{k} \left(j - \frac{k+1}{2} \right) \left(R_{ij} - \frac{(k+1)}{2} \right) \right\} \\ &= \frac{12}{\sqrt{n}\, k(k^2-1)} \sum_{i=1}^{n} r_i, \end{aligned}$$

r_i 为第 i 行 $(R_{i1}, \cdots, R_{ik})$ 与设定次序 $(1, \cdots, k)$ 的 Spearman 秩相关系数. 对于 Freidman 统计量 S 有

$$\begin{aligned} &\frac{1}{(n-1)(k-1)}[S-(k-1)] \\ &= \frac{12}{n(n-1)k(k^2-1)} \left\{ \sum_{t=1}^{k} \left[R_{\cdot t} - \frac{n(k+1)}{2} \right]^2 - \frac{nk(k^2-1)}{12} \right\} \\ &= \frac{12}{n(n-1)k(k^2-1)} \sum_{t=1}^{k} \left\{ \sum_{i \neq j}\sum \left(R_{it} - \frac{k+1}{2} \right) \left(R_{jt} - \frac{k+1}{2} \right) \right\} \\ &= \frac{2}{n(n-1)} \sum_{i<j}\sum \left\{ \frac{12}{k(k^2-1)} \sum_{t=1}^{k} \left(R_{it} - \frac{k+1}{2} \right) \left(R_{jt} - \frac{k+1}{2} \right) \right\} \end{aligned}$$

为任意两行间 Spearman 秩相关系数的平均.

对两个变量(X,Y)间考虑线性回归模型

$$Y_i=\beta_0+\beta X_i+\varepsilon_i,\quad i=1,\cdots,n,$$

$\varepsilon_1,\cdots,\varepsilon_n$ 相互独立同分布,分布连续. $X_1,\cdots,X_n$ 考虑为固定点,即为已知数,且假定 $X_1<\cdots<X_n$. 检验回归系数 β,假设

$$H_0: \beta=0,\qquad H_1: \beta>0.$$

问题的提法与正态回归问题类似. 在 ε_i 有正态分布 $N(0,\sigma^2)$时,检验上述假设的检验统计量为

$$\begin{aligned}T&=\frac{\hat\beta\sqrt{\sum(X_i-\overline{X})^2}}{\sqrt{\sum(Y_i-\hat\beta_0-\hat\beta X_i)^2}/\sqrt{n-2}}\\&=\frac{\sqrt{n-2}\cdot\left[\sum(Y_i-\overline{Y})(X_i-\overline{X})\right]}{\sqrt{\sum(X_i-\overline{X})^2}\sqrt{\sum(Y_i-\overline{Y})^2-\sum(Y_i-\overline{Y})(X_i-\overline{X})}}.\end{aligned}\tag{9.3.5}$$

在非参数情形,仅知 $\varepsilon_1,\cdots,\varepsilon_n$ 独立同分布,分布连续,检验上述假设,可以构造一个秩检验统计量. 在(9.3.5)式的 T 统计量中以$(X_1,\cdots,X_n)$的序$(1,\cdots,n)$代替$(X_1,\cdots,X_n)$,以$(Y_1,\cdots,Y_n)$的秩统计量$(R_1,\cdots,R_n)$代替$(Y_1,\cdots,Y_n)$,则得

$$T^*=\frac{\sqrt{n-2}\left[\sum\left(i-\frac{n+1}{2}\right)\left(R_i-\frac{n+1}{2}\right)\right]}{\sqrt{\sum\left(i-\frac{n+1}{2}\right)^2}\sqrt{\sum\left(R_i-\frac{n+1}{2}\right)^2-\sum\left(i-\frac{n+1}{2}\right)\left(R_i-\frac{n+1}{2}\right)}}.\tag{9.3.6}$$

T^* 取大的值有利于 H_1,否定域应为$\{T^*\geqslant t_\alpha\}$,$t_\alpha$ 为临界值. 这个否定域与相同显著性水平 α 的 Spearman 统计量的否定域

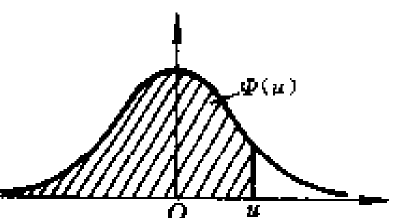
$$S=\frac{\sum\left(i-\frac{n+1}{2}\right)\left(R_i-\frac{n+1}{2}\right)}{\sqrt{\sum\left(i-\frac{n+1}{2}\right)^2\cdot\sum\left(R_i-\frac{n+1}{2}\right)^2}}\geqslant s_\alpha$$

是等价的. 也就是说可以用 Spearman 统计量检验回归系数 H_0: $\beta=0$.

附表

附表1 标准正态分布数值表

$$\Phi(u)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-\frac{x^2}{2}}dx \quad (u \geqslant 0)$$



u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	u
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	0.0
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753	0.1
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141	0.2
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517	0.3
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879	0.4
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224	0.5
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549	0.6
0.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7794	.7823	.7852	0.7
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133	0.8
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389	0.9
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621	1.0
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830	1.1
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.90147	1.2
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774	1.3
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189	1.4
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408	1.5
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449	1.6
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327	1.7
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062	1.8
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670	1.9
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.93124	.98169	2.0
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574	2.1
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899	2.2

续表 1

u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	u
2.3	.98928	.98956	.98988	$.9^{2}0097$	$.9^{2}0358$	$.9^{2}0613$	$.9^{2}0863$	$.9^{2}1106$	$.9^{2}1344$	$.9^{2}1576$	2.3
2.4	$.9^{2}1802$	$.9^{2}2024$	$.9^{2}2240$	$.9^{2}2451$	$.9^{2}2656$	$.9^{2}2857$	$.9^{2}3053$	$.9^{2}3244$	$.9^{2}3431$	$.9^{3}3613$	2.4
2.5	$.9^{2}3790$	$.9^{2}3963$	$.9^{2}4132$	$.9^{2}4297$	$.9^{2}4457$	$.9^{2}4614$	$.9^{2}4766$	$.9^{2}4915$	$.9^{2}5060$	$.9^{2}5201$	2.5
2.6	$.9^{2}5339$	$.9^{2}5473$	$.9^{2}5604$	$.9^{2}5731$	$.9^{2}5855$	$.9^{2}5975$	$.9^{2}6093$	$.9^{2}6207$	$.9^{2}6319$	$.9^{2}6427$	2.6
2.7	$.9^{2}6533$	$.9^{2}6636$	$.9^{2}6736$	$.9^{2}6833$	$.9^{2}6928$	$.9^{2}7020$	$.9^{2}7110$	$.9^{2}7197$	$.9^{2}7282$	$.9^{2}7365$	2.7
2.8	$.9^{2}7445$	$.9^{2}7523$	$.9^{2}7599$	$.9^{2}7673$	$.9^{2}7744$	$.9^{2}7814$	$.9^{2}7882$	$.9^{2}7948$	$.9^{2}8012$	$.9^{2}8074$	2.8
2.9	$.9^{2}8134$	$.9^{2}8193$	$.9^{2}8250$	$.9^{2}8305$	$.9^{2}8359$	$.9^{2}8411$	$.9^{2}8462$	$.9^{2}8511$	$.9^{2}8559$	$.9^{2}8605$	2.9
3.0	$.9^{2}8650$	$.9^{2}8694$	$.9^{2}8736$	$.9^{2}8777$	$.9^{2}8817$	$.9^{2}8856$	$.9^{2}8893$	$.9^{2}8930$	$.9^{2}8965$	$.9^{2}8999$	3.0
3.1	$.9^{3}0324$	$.9^{3}0646$	$.9^{3}0957$	$.9^{3}1260$	$.9^{3}1553$	$.9^{3}1836$	$.9^{3}2112$	$.9^{3}2378$	$.9^{3}2636$	$.9^{3}2886$	3.1
3.2	$.9^{3}3129$	$.9^{3}3363$	$.9^{3}3590$	$.9^{3}3810$	$.9^{3}4024$	$.9^{3}4230$	$.9^{3}4429$	$.9^{3}4623$	$.9^{3}4810$	$.9^{3}4991$	3.2
3.3	$.9^{3}5166$	$.9^{3}5335$	$.9^{3}5499$	$.9^{3}5658$	$.9^{3}5811$	$.9^{3}5959$	$.9^{3}6103$	$.9^{3}6242$	$.9^{3}6376$	$.9^{3}6505$	3.3
3.4	$.9^{3}6631$	$.9^{3}6752$	$.9^{3}6869$	$.9^{3}6982$	$.9^{3}7091$	$.9^{3}7197$	$.9^{3}7299$	$.9^{3}7398$	$.9^{3}7493$	$.9^{3}7585$	3.4
3.5	$.9^{3}7674$	$.9^{3}7759$	$.9^{3}7842$	$.9^{3}7922$	$.9^{3}7999$	$.9^{3}8074$	$.9^{3}8146$	$.9^{3}8215$	$.9^{3}8282$	$.9^{3}8347$	3.5
3.6	$.9^{3}8409$	$.9^{3}8469$	$.9^{3}8527$	$.9^{3}8583$	$.9^{3}8637$	$.9^{3}8689$	$.9^{3}8739$	$.9^{3}8787$	$.9^{3}8834$	$.9^{3}8879$	3.6
3.7	$.9^{3}8922$	$.9^{3}8964$	$.9^{4}0039$	$.9^{4}0426$	$.9^{4}0799$	$.9^{4}1158$	$.9^{4}1504$	$.9^{4}1838$	$.9^{4}2159$	$.9^{4}2468$	3.7
3.8	$.9^{4}2765$	$.9^{4}3052$	$.9^{4}3327$	$.9^{4}3593$	$.9^{4}3848$	$.9^{4}4094$	$.9^{4}4331$	$.9^{4}4558$	$.9^{4}4777$	$.9^{4}4988$	3.8
3.9	$.9^{4}5190$	$.9^{4}5385$	$.9^{4}5573$	$.9^{4}5753$	$.9^{4}5926$	$.9^{4}6092$	$.9^{4}6253$	$.9^{4}6406$	$.9^{4}6554$	$.9^{4}6696$	3.9
4.0	$.9^{4}6833$	$.9^{4}6964$	$.9^{4}7090$	$.9^{4}7211$	$.9^{4}7327$	$.9^{4}7439$	$.9^{4}7546$	$.9^{4}7649$	$.9^{4}7748$	$.9^{4}7843$	4.0
4.1	$.9^{4}7934$	$.9^{4}8022$	$.9^{4}8106$	$.9^{4}8186$	$.9^{4}8263$	$.9^{4}8338$	$.9^{4}8409$	$.9^{4}8477$	$.9^{4}8542$	$.9^{4}8605$	4.1
4.2	$.9^{4}8665$	$.9^{4}8723$	$.9^{4}8778$	$.9^{4}8832$	$.9^{4}8882$	$.9^{4}8931$	$.9^{4}8978$	$.9^{5}0226$	$.9^{5}0655$	$.9^{5}1066$	4.2
4.3	$.9^{5}1460$	$.9^{5}1837$	$.9^{5}2199$	$.9^{5}2545$	$.9^{5}2876$	$.9^{5}3193$	$.9^{5}3497$	$.9^{5}3788$	$.9^{5}4066$	$.9^{5}4332$	4.3
4.4	$.9^{5}4587$	$.9^{5}4831$	$.9^{5}5065$	$.9^{5}5288$	$.9^{5}5502$	$.9^{5}5706$	$.9^{5}5902$	$.9^{5}6089$	$.9^{5}6268$	$.9^{5}6439$	4.4
4.5	$.9^{5}6602$	$.9^{5}6759$	$.9^{5}6908$	$.9^{5}7051$	$.9^{5}7187$	$.9^{5}7318$	$.9^{5}7442$	$.9^{5}7561$	$.9^{5}7675$	$.9^{5}7784$	4.5
4.6	$.9^{5}7888$	$.9^{5}7987$	$.9^{5}8081$	$.9^{5}8172$	$.9^{5}8258$	$.9^{5}8340$	$.9^{5}8419$	$.9^{5}8494$	$.9^{5}8566$	$.9^{5}8634$	4.6
4.7	$.9^{5}8699$	$.9^{5}8761$	$.9^{5}8821$	$.9^{5}8877$	$.9^{5}8931$	$.9^{5}8983$	$.9^{6}0320$	$.9^{6}0789$	$.9^{6}1235$	$.9^{6}1661$	4.7
4.8	$.9^{6}2067$	$.9^{6}2453$	$.9^{6}2822$	$.9^{6}3173$	$.9^{6}3508$	$.9^{6}3827$	$.9^{6}4131$	$.9^{6}4420$	$.9^{6}4696$	$.9^{6}4958$	4.8
4.9	$.9^{6}5208$	$.9^{6}5446$	$.9^{6}5673$	$.9^{6}5889$	$.9^{6}6094$	$.9^{6}6289$	$.9^{6}6475$	$.9^{6}6652$	$.9^{6}6821$	$.9^{6}6981$	4.9

附表2 χ^2 分布临界值表

n \ λ \ α	0.975	0.05	0.025	0.01
1	0.00098	3.84	5.02	6.63
2	0.0506	5.99	7.38	9.21
3	0.216	7.81	9.35	11.3
4	0.484	9.49	11.1	13.3
5	0.831	11.07	12.8	15.1
6	1.24	12.6	14.4	16.8
7	1.69	14.1	16.0	18.5
8	2.18	15.5	17.5	20.1
9	2.70	16.9	19.0	21.7
10	3.25	18.3	20.5	23.2
11	3.82	19.7	21.9	24.7
12	4.40	21.0	23.3	26.2
13	5.01	22.4	24.7	27.7
14	5.63	23.7	26.1	29.1
15	6.26	25.0	27.5	30.6
16	6.91	26.3	28.8	32.0
17	7.56	27.6	30.2	33.4
18	8.23	28.9	31.5	34.8
19	8.91	30.1	32.9	36.2
20	9.59	31.4	34.2	37.6
21	10.3	32.7	35.5	38.9
22	11.0	33.9	36.8	40.3
23	11.7	35.2	38.1	41.6
24	12.4	36.4	39.4	43.0
25	13.1	37.7	40.6	44.3
26	13.8	38.9	41.9	45.6
27	14.6	40.1	43.2	47.0
28	15.3	41.3	44.5	48.3
29	16.0	42.6	45.7	49.6
30	16.8	43.8	47.0	50.9

［注］ n：自由度，λ：临界值，$P\{\chi^2>\lambda\}=\alpha$.

附表 3　符号检验临界值表

（本表列出了满足 $P\{B \geqslant b\} \leqslant \alpha$ 的临界值 b）

n \ α	0.01	0.05	0.10	n \ α	0.01	0.05	0.10	n \ α	0.01	0.05	0.10
4			4	17	14	13	12	34	25	23	22
5		5	5	18	15	13	13	35	25	23	22
6		6	6	19	15	14	13	36	26	24	23
7	7	7	6	20	16	15	14	37	26	24	23
8	8	7	7	21	17	15	14	38	27	25	24
9	9	8	7	22	17	16	15	39	28	26	24
10	10	9	8	23	18	16	16	40	28	26	25
11	10	9	9	24	19	17	16	41	29	27	26
12	11	10	9	25	19	18	17	42	29	27	26
13	12	10	10	26	20	18	17	43	30	28	27
14	12	11	10	27	20	19	18	44	31	28	27
15	13	12	11	28	21	19	18	45	31	29	28
16	14	12	12	29	22	20	19	46	32	30	28
				30	22	20	20	47	32	30	29
				31	23	21	20	48	33	31	29
				32	24	22	21	49	34	31	30
				33	24	22	21	50	34	32	31

附表 4 Wilcoxon 两样本秩和检验临界值表

（本表列出满足 $P\{W\leqslant c_1\}\leqslant\frac{\alpha}{2}$，$P\{W\geqslant c_2\}\leqslant\frac{\alpha}{2}$ 的临界值 c_1，c_2）

n_1	n_2	$\alpha=0.10$		$\alpha=0.05$	
		c_1	c_2	c_1	c_2
2	2	3	7	3	7
	3	3	9	3	9
	4	3	11	3	11
	5	4	12	3	13
	6	4	14	3	15
	7	4	16	3	17
	8	5	17	4	18
	9	5	19	4	20
	10	5	21	4	22
	11	5	23	4	24
	12	6	24	5	25
	13	6	26	5	27
	14	7	27	5	29
	15	7	29	5	31
	16	7	31	5	33
	17	7	33	6	34
	18	8	34	6	36
	19	8	36	6	38
	20	8	38	6	40
3	3	7	14	6	15
	4	7	17	6	18
	5	8	19	7	20
	6	9	21	8	22
	7	9	24	8	25
	8	10	26	9	27
	9	11	28	9	30
	10	11	31	10	32
	11	12	33	10	35
	12	12	36	11	37
	13	13	38	11	40
	14	14	40	12	42
	15	14	43	12	45
	16	15	45	13	47
	17	16	47	13	50
	18	16	50	14	52
	19	17	52	14	55
	20	18	54	15	57

n_1	n_2	$\alpha=0.10$		$\alpha=0.05$	
		c_1	c_2	c_1	c_2
4	4	12	24	11	25
	5	13	27	12	28
	6	14	30	13	31
	7	15	33	14	34
	8	16	36	15	37
	9	17	39	15	41
	10	18	42	16	44
	11	19	45	17	47
	12	20	48	18	50
	13	21	51	19	53
	14	22	54	20	56
	15	23	57	21	59
	16	25	59	22	62
	17	26	62	22	66
	18	27	65	23	69
	19	28	68	24	72
	20	29	71	25	75
5	5	20	35	18	37
	6	21	39	19	41
	7	22	43	21	44
	8	24	46	22	48
	9	25	50	23	52
	10	27	53	24	56
	11	28	57	25	60
	12	29	61	27	63
	13	31	64	28	67
	14	32	68	29	71
	15	34	71	30	75
	16	35	75	31	79
	17	36	79	33	82
	18	38	82	34	86
	19	39	86	35	90
	20	41	89	36	94
6	6	29	49	27	51
	7	30	54	28	56
	8	32	58	30	60
	9	34	62	32	64

续表 4

n_1	n_2	$\alpha=0.10$		$\alpha=0.05$	
		c_1	c_2	c_1	c_2
6	10	36	66	33	69
	11	38	70	35	73
	12	39	75	36	78
	13	41	79	38	82
	14	43	83	39	87
	15	45	87	41	91
	16	47	91	43	95
	17	48	96	44	100
	18	50	100	46	104
	19	52	104	47	109
	20	54	108	49	113
7	7	40	65	37	68
	8	42	70	39	73
	9	44	75	41	78
	10	46	80	43	83
	11	48	85	45	88
	12	50	90	47	93
	13	53	94	49	98
	14	55	99	51	103
	15	57	104	53	108
	16	59	109	55	113
	17	62	113	57	118
	18	64	118	59	123
	19	66	123	61	128
	20	68	128	63	133
8	8	52	84	50	86
	9	55	89	52	92
	10	57	95	54	98
	11	60	100	56	104
	12	63	105	59	109
	13	65	111	61	115
	14	68	116	63	121
	15	70	122	66	126
	16	73	127	68	132
	17	76	132	71	137
	18	78	138	73	143
	19	81	143	75	149
	20	84	148	78	154
9	9	67	104	63	108
	10	70	110	66	114
	11	73	116	69	120
	12	76	122	72	126
	13	79	128	74	133
	14	82	134	77	139
	15	85	140	80	145
	16	88	146	83	151
	17	91	152	85	158
	18	94	158	88	164
	19	97	164	91	170
	20	100	170	94	176
10	10	83	127	79	131
	11	87	133	82	138
	12	90	140	85	145
	13	93	147	89	151
	14	97	153	92	158
	15	100	160	95	165
	16	104	166	98	172
	17	107	173	101	179
	18	111	179	104	186
	19	114	186	108	192
	20	118	192	111	199
11	11	101	152	97	156
	12	105	159	100	164
	13	109	166	104	171
	14	113	173	107	179
	15	117	180	111	186
	16	121	187	114	194
	17	124	195	118	201
	18	128	202	122	208
	19	132	209	125	216
	20	136	216	129	223
12	12	121	179	116	184
	13	126	186	120	192
	14	130	194	124	200
	15	134	202	128	208
	16	139	209	132	216
	17	143	217	136	224
	18	147	225	140	232
	19	151	233	144	240
	20	156	240	148	248

续表 4

n_1	n_2	$\alpha=0.10$		$\alpha=0.05$		n_1	n_2	$\alpha=0.10$		$\alpha=0.05$	
		c_1	c_2	c_1	c_2			c_1	c_2	c_1	c_2
13	13	143	208	137	214	15	20	221	319	211	329
	14	148	216	142	222						
	15	153	224	146	231	16	16	220	308	212	316
	16	157	233	151	239		17	226	318	218	326
	17	162	241	155	248		18	232	328	223	337
	18	167	249	159	257		19	238	338	229	347
	19	172	257	164	265		20	244	348	235	357
	20	176	266	168	274						
						17	17	250	345	241	354
14	14	167	239	161	245		18	256	356	247	365
	15	172	248	165	255		19	263	366	253	376
	16	177	257	170	264		20	269	377	259	387
	17	183	265	175	273						
	18	188	274	180	282	18	18	281	385	271	395
	19	193	283	184	292		19	288	396	278	406
	20	198	292	189	301		20	295	407	284	418
15	15	193	272	185	280	19	19	314	427	304	437
	16	198	282	191	289		20	321	439	310	450
	17	204	291	196	299						
	18	209	301	201	309	20	20	349	471	338	482
	19	215	310	206	319						

[注] 当第二样本量 n 小于或等于第一样本量 m 时，取 $n_1=n, n_2=m$，从表中可直接查出临界值 c_1, c_2。

当 $n>m$ 时，取 $n_1=m, n_2=n$，从表中查出 c_1, c_2，此时检验的临界值为

$$c_1' = \frac{1}{2}(m+n)(m+n+1) - c_2,$$

$$c_2' = \frac{1}{2}(m+n)(m+n+1) - c_1$$

附表 5 Wilcoxon 符号秩检验临界值表

(本表列出满足 $P\{W^+\leqslant c_1\}\leqslant\frac{\alpha}{2}, P\{W^+\geqslant c_2\}\leqslant\frac{\alpha}{2}$ 的临界值 c_1, c_2)

n	$\alpha=0.10$		$\alpha=0.05$		n	$\alpha=0.10$		$\alpha=0.05$	
	c_1	c_2	c_1	c_2		c_1	c_2	c_1	c_2
5	0	15	—	—	18	47	124	40	131
6	2	19	0	21	19	53	137	46	144
7	3	25	2	26	20	60	150	52	158
8	5	31	3	33	21	67	164	58	173
9	8	37	5	40	22	75	178	66	187
10	10	45	8	47	23	83	193	73	203
11	13	53	10	56	24	91	209	81	219
12	17	61	13	65	25	100	225	89	236
13	21	70	17	74	26	110	241	98	253
14	25	80	21	84	27	119	259	107	271
15	30	90	25	95	28	130	276	116	290
16	35	101	29	107	29	140	295	126	309
17	41	112	34	119	30	151	314	137	328

附表 6 Kruskal-Wallis k 样本检验统计量 H^* 临界值表

（本表列出对应不同样本量满足 $P\{H^* \geq h_\alpha\} = \alpha$ 的临界值 h_α 及对应的 α.）

样本量			临界值	α	样本量			临界值	α
n_1	n_2	n_3	h_α		n_1	n_2	n_3	h_α	
2	1	1	2.7000	0.500	4	3	2	6.4444	0.008
2	2	1	3.6000	0.200				6.3000	0.011
2	2	2	4.5714	0.067				5.4444	0.046
			3.7143	0.200				5.4000	0.051
3	1	1	3.2000	0.300				4.5111	0.098
3	2	1	4.2857	0.100				4.4444	0.102
			3.8571	0.133	4	3	3	6.7455	0.010
3	2	2	5.3572	0.029				6.7091	0.013
			4.7143	0.048				5.7909	0.046
			4.5000	0.067				5.7273	0.050
			4.4643	0.105				4.7091	0.092
3	3	1	5.1429	0.043				4.7000	0.101
			4.5714	0.100	4	4	1	6.6667	0.010
			4.0000	0.129				6.1667	0.022
3	3	2	6.2500	0.011				4.9667	0.048
			5.3611	0.032				4.8667	0.054
			5.1389	0.061				4.1667	0.082
			4.5556	0.100				4.0667	0.102
			4.2500	0.121	4	4	2	7.0364	0.006
3	3	3	7.2000	0.004				6.8727	0.011
			6.4889	0.011				5.4545	0.046
			5.6889	0.029				5.2364	0.052
			5.6000	0.050				4.5545	0.098
			5.0667	0.086				4.4455	0.103
			4.6222	0.100	4	4	3	7.1439	0.010
4	1	1	3.5714	0.200				7.1364	0.011
4	2	1	4.8214	0.057				5.5985	0.049
			4.5000	0.076				5.5758	0.051
			4.0179	0.114				4.5455	0.099
4	2	2	6.0000	0.014				4.4773	0.102
			5.3333	0.033	4	4	4	7.6538	0.008
			5.1250	0.052				7.5385	0.011
			4.4583	0.100				5.6923	0.049
			4.1667	0.105				5.6538	0.054
4	3	1	5.8333	0.021				4.6539	0.097
			5.2083	0.050				4.5001	0.104
			5.0000	0.057	5	1	1	3.8571	0.143
			4.0556	0.093	5	2	1	5.2500	0.036
			3.8889	0.129				5.0000	0.048

续表 6

样本量			临界值 h_c	α
n_1	n_2	n_3		
5	2	1	4.4500	0.071
			4.2000	0.095
			4.0500	0.119
5	2	2	6.5333	0.008
			6.1333	0.013
			5.1600	0.034
			5.0400	0.056
			4.3733	0.090
			4.2933	0.122
5	3	1	6.4000	0.012
			4.9600	0.048
			4.8711	0.052
			4.0178	0.095
			3.8400	0.123
5	3	2	6.9091	0.009
			6.8218	0.010
			5.2509	0.049
			5.1055	0.052
			4.6509	0.091
			4.4945	0.101
5	3	3	7.0788	0.009
			6.9818	0.011
			5.6485	0.049
			5.5152	0.051
			4.5333	0.097
			4.4121	0.109
5	4	1	6.9545	0.008
			6.8400	0.011
			4.9855	0.044
			4.8600	0.056
			3.9873	0.098
			3.9600	0.102
5	4	2	7.2045	0.009
			7.1182	0.010
			5.2727	0.049
			5.2682	0.050
			4.5409	0.098
			4.5182	0.101
5	4	3	7.4449	0.010
			7.3949	0.011
			5.6564	0.049
			5.6308	0.050
			4.5487	0.099
			4.5231	0.103
5	4	4	7.7604	0.009
			7.7440	0.011
			5.6571	0.049
			5.6176	0.050
			4.6187	0.100
			4.5527	0.102
5	5	1	7.3091	0.009
			6.8364	0.011
			5.1273	0.046
			4.9091	0.053
			4.1091	0.086
			4.0364	0.105
5	5	2	7.3385	0.010
			7.2692	0.010
			5.3385	0.047
			5.2462	0.051
			4.6231	0.097
			4.5077	0.100
5	5	3	7.5780	0.010
			7.5429	0.010
			5.7055	0.046
			5.6264	0.051
			4.5451	0.100
			4.5363	0.102
5	5	4	7.8229	0.010
			7.7914	0.010
			5.6657	0.049
			5.6429	0.050
			4.5229	0.099
			4.5200	0.101
5	5	5	8.0000	0.009
			7.9800	0.010
			5.7800	0.049
			5.6600	0.051
			4.5600	0.100
			4.5000	0.102

附表 7(A)　Jonckheere-Terpstra k 样本检验统计量 J 的临界值表(A)

(本表列出邻近 $P\{J\geqslant k_\alpha\}=\alpha$ 的临界值 k_α. 括号中的值是该值对应的真实的显著性水平.)

样本量						样本量					
n_1	n_2	n_3	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.025$	n_1	n_2	n_3	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.025$
2	2	2	9(.16667)	10(.08889)	11(0.3333)	2	3	7	29(.10960)	31(.06023)	33(.02929)
			10(.08889)	11(.03333)	12(.01111)				30(.08232)	32(.04268)	34(.01032)
2	2	3	12(.13810)	13(.07619)	14(.03810)	2	3	8	32(.11826)	35(.05198)	37(.02650)
			13(.07619)	14(.03810)	15(.01429)				33(.09192)	36(.03768)	38(.01810)
2	2	4	15(.11667)	16(.07143)	17(.03810)	2	4	4	23(.10794)	25(.05016)	26(.03206)
			16(.07143)	17(.03810)	18(.01905)				24(.07556)	26(.03206)	27(.01905)
2	2	5	18(.10450)	19(.06614)	20(.03968)	2	4	5	27(.10491)	29(.05397)	30(.03680)
			19(.06614)	20(.03968)	21(.02116)				28(.07662)	30(.03680)	31(.02395)
2	2	6	20(.13571)	22(.06349)	23(.03968)	2	4	6	31(.10245)	33(.05685)	35(.02821)
			21(.09444)	23(.03968)	24(.02381)				32(.07742)	34(.04076)	36(.01898)
2	2	7	23(.12172)	25(.06061)	27(.02525)	2	4	7	35(.10047)	37(.05921)	39(.03193)
			24(.08788)	26(.04040)	28(.01515)				36(.07797)	38(.04406)	40(.02261)
2	2	8	26(.11178)	28(.05892)	30(.02694)	2	4	8	38(.12266)	41(.06112)	44(.02593)
			27(.08215)	29(.04040)	31(.01684)				39(.09879)	42(.04686)	45(.01863)
2	3	3	15(.15179)	17(.05714)	18(.03036)	2	5	5	31(.11935)	34(.05014)	35(.03565)
			16(.09643)	18(.03036)	19(.01429)				32(.09157)	35(.03565)	36(.02453)
2	3	4	19(.11190)	20(.07381)	22(.02619)	2	5	6	36(.10462)	38(.06277)	40(.03469)
			20(.07381)	21(.04524)	23(.01349)				37(.08178)	39(.04715)	41(.02486)
2	3	5	22(.12421)	24(.05913)	25(.03810)	2	5	7	40(.11588)	43(.05821)	46(.02507)
			23(.08770)	25(.03810)	26(.02302)				41(.09355)	44(.04477)	47(.01820)
2	3	6	25(.13398)	27(.07143)	29(.03290)	2	5	8	45(.10400)	48(.05459)	51(.02519)
			26(.09957)	28(.04957)	30(.02100)				46(.08500)	49(.04283)	52(.01885)

续表 7(A)

样本量					
n_1	n_2	n_3	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.025$
2	6	6	41(.10607)	44(.05260)	46(.03031)
			42(.08528)	45(.04027)	47(.02235)
2	6	7	46(.10721)	49(.05720)	52(.02703)
			47(.08803)	50(.04521)	53(.02040)
2	6	8	51(.10804)	54(.06118)	57(.03135)
			52(.09031)	55(.04953)	58(.02449)
2	7	7	52(.10029)	55(.05628)	58(.02858)
			53(.08358)	56(.04543)	59(0.2225)
2	7	8	57(.11128)	61(.05543)	64(.02987)
			58(.09468)	62(.04555)	65(.02381)
2	8	8	63(.11392)	68(.05085)	71(.02858)
			64(.09833)	69(.04231)	72(.02319)
3	3	3	19(.13869)	21(.06131)	22(.03690)
			20(.09464)	22(.03690)	23(.02083)
3	3	4	23(.13000)	25(.06405)	27(.02643)
			24(.09310)	26(.04214)	28(.01548)
3	3	5	27(.12348)	29(.06623)	31(.03106)
			28(.09177)	30(.04621)	32(.02002)
3	3	6	31(.11845)	33(.06791)	35(.03506)
			32(.09075)	34(.04946)	36(.02408)
3	3	7	35(.11451)	38(.05219)	40(.02768)
			36(.08989)	39(.03849)	41(.01941)
3	3	8	39(.11131)	42(.05446)	44(.03092)
			40(.08914)	43(.04144)	45(.02259)

样本量					
n_1	n_2	n_3	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.025$
3	4	4	28(.10926)	30(.05758)	32(.02649)
			29(.08043)	31(.03974)	33(.01688)
3	4	5	32(.12269)	35(.05281)	37(.02648)
			33(.09481)	36(.03791)	38(.01789)
3	4	6	37(.10723)	39(.06505)	42(.02642)
			38(.08432)	40(.04923)	43(.01865)
3	4	7	41(.11810)	44(.06003)	47(.02633)
			42(.09566)	45(.04644)	48(.01926)
3	4	8	46(.10583)	49(.05607)	52(.02624)
			47(.08672)	50(.04419)	53(.01974)
3	5	5	37(.12200)	40(.05823)	42(.03227)
			38(.09706)	41(.04382)	43(.02324)
3	5	6	42(.12137)	45(.06278)	48(.02822)
			43(.09882)	46(.04890)	49(.02085)
3	5	7	48(.10022)	51(.05332)	54(.02518)
			49(.08220)	52(.04211)	55(.01903)
3	5	8	53(.10138)	56(.05718)	59(.02926)
			54(.08461)	57(.04627)	60(.02284)
3	6	6	48(.11162)	51(.06089)	54(.02965)
			49(.09226)	52(.04855)	55(.02267)
3	6	7	54(.10392)	57(.05931)	60(.03081)
			55(.08704)	58(.04821)	61(.02420)
3	6	8	59(.11440)	63(.05797)	67(.02551)
			60(.09770)	64(.04788)	68(.02027)

续表 7(A)

样本量						样本量					
n_1	n_2	n_3	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.025$	n_1	n_2	n_3	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.025$
3	7	7	60(.10697)	64(.05366)	67(.02919)	4	6	6	55(.11612)	59(.05592)	62(.02909)
			61(.09112)	65(.04421)	68(.02337)				56(.09810)	60(.04546)	63(.02287)
3	7	8	66(.10953)	70(.05853)	74(.02783)	4	6	7	62(.10126)	66(.05067)	69(.02756)
			67(.09460)	71(.04917)	75(.02265)				63(.08619)	67(.04174)	70(.02208)
3	8	8	73(.10544)	78(.05022)	81(.02986)	4	6	8	68(.10397)	72(.05539)	76(.02631)
			74(.09197)	79(.04251)	82(.02477)				69(.08972)	73(.04651)	77(.02141)
4	4	4	33(.10993)	35(.06323)	37(.03296)	4	7	7	68(.11261)	73(.05154)	76(.02963)
			34(.08439)	36(.04632)	38(.02286)				69(.09759)	74(.04318)	77(.02426)
4	4	5	38(.11051)	41(.05178)	43(.02833)	4	7	8	75(.10806)	80(.05226)	84(.02621)
			39(.08738)	42(.03873)	44(.02027)				76(.09450)	81(.04441)	85(.02170)
4	4	6	43(.11087)	46(.05649)	48(.03336)	4	8	8	82(.11160)	87(.05754)	92(.02610)
			44(.08984)	47(.04376)	49(.02497)				83(.09869)	88(.04966)	93(.02191)
4	4	7	48(.11118)	51(.06052)	54(.02939)	5	5	5	50(.10490)	53(.05715)	56(.02788)
			49(.09184)	52(.04822)	55(.02244)				51(.08666)	54(.04558)	57(.02136)
4	4	8	53(.11139)	57(.05216)	60(.02636)	5	5	6	56(.10781)	60(.05124)	63(.02637)
			54(.09353)	58(.04204)	61(.02049)				57(.09078)	61(.04151)	64(.02066)
4	5	5	44(.10139)	47(.05094)	49(.02980)	5	5	7	62(.11026)	66(.05631)	70(.02514)
			45(.08177)	48(.03928)	50(.02220)				63(.09430)	67(.04665)	71(.02008)
4	5	6	49(.11377)	53(.05021)	55(.03096)	5	5	8	68(.11235)	73(.05135)	76(.02948)
			50(.09435)	54(.03970)	56(.02382)				69(.09734)	74(.04300)	77(.02413)
4	5	7	55(.10570)	58(.06081)	62(.02519)	5	6	6	63(.10301)	67(.05205)	70(.02859)
			56(.08875)	59(.04959)	63(.01963)				64(.08787)	68(.04301)	71(.02299)
4	5	8	60(.11594)	64(.05923)	68(.02636)	5	6	7	69(.11412)	74(.05272)	78(.02507)
			61(.09919)	65(.04905)	69(.02102)				70(.09906)	75(.04427)	79(.02042)

续表 7(A)

样本量						样本量					
n_1	n_2	n_3	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.025$	n_1	n_2	n_3	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.025$
5	6	8	76(.10935)	81(.05328)	85(.02694)	6	7	7	85(.10808)	90(.05595)	95(.02559)
			77(.09575)	82(.04535)	86(.02235)				86(.09563)	91(.04835)	96(.02153)
5	7	7	77(.10319)	81(.05828)	85(.02998)	6	7	8	93(.10571)	99(.05002)	103(.02787)
			78(.09019)	82(.04981)	86(.02499)				94(.09426)	100(.04351)	104(.02380)
5	7	8	84(.10680)	89(.05493)	93(.02948)	6	8	8	101(.10987)	107(.05532)	112(.02822)
			85(.09438)	90(.04739)	94(.02489)				102(.09885)	108(.04873)	113(.02437)
5	8	8	92(.10461)	97(.05635)	102(.02724)	7	7	7	94(.10045)	99(.05401)	104(.02609)
			93(.09319)	98(.04917)	103(.02322)				95(.08944)	100(.04711)	105(.02225)
6	6	6	70(.10721)	74(.05805)	78(.02816)	7	7	8	102(.10477)	108(.05235)	113(.02653)
			71(.09285)	75(.04897)	79(.02306)				103(.09414)	109(.04605)	114(.02288)
6	6	7	77(.11084)	82(.05446)	86(.02778)	7	8	8	111(.10393)	117(.05443)	123(.02539)
			78(.09721)	83(.04645)	87(.02311)				112(.09402)	118(.04834)	124(.02209)
6	6	8	85(.10104)	90(.05151)	94(.02745)	8	8	8	120(.10852)	127(.05365)	133(.02629)
			86(.08914)	91(.04436)	95(.02313)				121(.09891)	128(.04798)	134(.02310)

附表 7(B)　Jonckheere-Terpstra k 样本检验统计量 J 的临界值表(B)

(本表列出当 k 组样本样本量 $n_1=n_2=\cdots=n_k=n$ 时，邻近 $P\{J\geqslant k_\alpha\}=\alpha$ 的临界值 k_α. 括号中的值是该值对应的真实的显著性水平.)

	k	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.025$
$n=2$	4	17(.13016) 18(.08294)	18(.08294) 19(.04841)	20(.02619) 21(.01230)
	5	27(.12133) 28(.08779)	29(.06126) 30(.04116)	31(.02646) 32(.01623)
	6	39(.12151) 40(.09533)	42(.05533) 43(.04083)	44(.02944) 45(.02071)
$n=3$	4	36(.11663) 37(.09067)	39(.05145) 40(.03744)	41(.02657) 42(.01834)
	5	58(.10487) 59(.08738)	61(.05884) 62(.04752)	64(.02995) 65(.02335)
	6	84(.11087) 85(.09686)	89(.05331) 90(.04524)	93(.02622) 94(.02201)
$n=4$	4	62(.10581) 63(.08950)	66(.05142) 67(.04198)	69(.02715) 70(.02150)
	5	99(.11129) 100(.09910)	105(.05211) 106(.04523)	109(.02876) 110(.02450)
	6	146(.10048) 147(.09181)	153(.05084) 154(.04567)	159(.02572) 160(.02274)
$n=5$	4	94(.10832) 95(.09621)	99(.05735) 100(.04983)	104(.02708) 105(.02296)
	5	152(.10385) 153(.09542)	159(.05492) 160(.04970)	166(.02603) 167(.02318)
	6	223(.10319) 224(.09679)	233(.05153) 234(.04775)	241(.02701) 242(.02477)
$n=6$	4	133(.10521) 134(.09607)	140(.05287) 141(.04743)	146(.02647) 147(.02336)
	5	215(.10494) 216(.09842)	225(.05229) 226(.04844)	234(.02505) 235(.02292)
	6	316(.10478) 317(.09982)	329(.05285) 330(.04990)	341(.02523) 342(.02361)

附表 8(A)　Friedman 二因素分组检验统计量 S 的临界值表(A)

(本表列出当 $\alpha=0.05$ 时,满足 $P\{S\geqslant s_\alpha\}\leqslant\alpha$ 的临界值 s_α. 表的左上角是由 S 的精确分布确定的值,右下角是由近似 χ^2 分布确定的值. 表中 k 是处理数,n 是区组数.)

n \ k	3	4	5	6	7	8	9	10	11	12	13	14	15
3	6.000	7.40	8.53	9.86	11.24	12.57	13.88	15.19	16.48	17.76	19.02	20.27	21.53
4	6.500	7.80	8.80	10.24	11.63	12.99	14.34	15.67	16.98	18.30	19.60	20.90	22.10
5	6.400	7.80	8.99	10.43	11.84	13.23	14.59	15.93	17.27	18.60	19.90	21.20	22.40
6	7.000	7.60	9.08	10.54	11.97	13.38	14.76	16.12	17.40	18.80	20.10	21.40	22.70
7	7.143	7.80	9.11	10.62	12.07	13.48	14.87	16.23	17.60	18.90	20.20	21.50	22.80
8	6.250	7.65	9.19	10.68	12.14	13.56	14.95	16.32	17.70	19.00	20.30	21.60	22.90
9	6.222	7.66	9.22	10.73	12.19	13.61	15.02	16.40	17.70	19.10	20.40	21.70	23.00
10	6.200	7.67	9.25	10.76	12.23	13.66	15.07	16.44	17.80	19.20	20.50	21.80	23.10
11	6.545	7.68	9.27	10.79	12.27	13.70	15.11	16.48	17.90	19.20	20.50	21.80	23.10
12	6.167	7.70	9.29	10.81	12.29	13.73	15.15	16.53	17.90	19.30	20.60	21.90	23.20
13	6.000	7.70	9.30	10.83	12.32	13.76	15.17	16.56	17.90	19.30	20.60	21.90	23.20
14	6.143	7.71	9.32	10.85	12.34	13.78	15.19	16.58	17.90	19.30	20.60	21.90	23.20
15	6.400	7.72	9.33	10.87	12.35	13.80	15.20	16.60	18.00	19.30	20.60	21.90	23.20
16	5.99	7.73	9.34	10.88	12.37	13.81	15.23	16.60	18.00	19.30	20.70	22.00	23.20
17	5.99	7.73	9.34	10.89	12.38	13.83	15.20	16.60	18.00	19.30	20.70	22.00	23.30
18	5.99	7.73	9.36	10.90	12.39	13.83	15.20	16.60	18.00	19.40	20.70	22.00	23.30
19	5.99	7.74	9.36	10.91	12.40	13.80	15.30	16.70	18.00	19.40	20.70	22.00	23.30
20	5.99	7.74	9.37	10.92	12.41	13.80	15.30	16.70	18.00	19.40	20.70	22.00	23.30
∞	5.99	7.82	9.49	11.07	12.59	14.07	15.51	16.92	18.31	19.68	21.03	22.36	23.69

附表 8(B)　Friedman 二因素分组检验统计量 S 的临界值表(B)

(本表列出当 $\alpha=0.01$ 时,满足 $P\{S\geqslant s_\alpha\}\leqslant\alpha$ 的临界值 s_α. 表的左上角是由 S 的精确分布确定的值,右下角是由近似 χ^2 分布确定的值. 表中 k 是处理数,n 是区组数.)

b \ k	3	4	5	6	7	8	9	10	11	12	13	14	15
3	—	9.000	10.13	11.76	13.26	14.78	16.28	17.74	19.19	20.61	22.00	23.38	24.76
4	8.000	9.600	11.20	12.59	14.19	15.75	17.28	18.77	20.24	21.70	23.10	24.50	25.90
5	8.400	9.960	11.43	13.11	14.74	16.32	17.86	19.37	20.86	22.30	23.70	25.20	26.60
6	9.000	10.200	11.75	13.45	15.10	16.69	18.25	19.77	21.30	22.70	24.20	25.60	27.00
7	8.857	10.371	11.97	13.69	15.35	16.95	18.51	20.04	21.50	23.00	24.40	25.90	27.30
8	9.000	10.350	12.14	13.87	15.53	17.15	18.71	20.24	21.80	23.20	24.70	26.10	27.50
9	8.667	10.44	12.27	14.01	15.68	17.29	18.87	20.42	21.90	23.40	24.95	26.30	27.70
10	9.600	10.53	12.38	14.12	15.79	17.41	19.00	20.53	22.00	23.50	25.00	26.40	27.90
11	9.455	10.60	12.46	14.21	15.89	17.52	19.10	20.64	22.10	23.60	25.10	26.60	28.00
12	9.500	10.68	12.53	14.28	15.96	17.59	19.19	20.73	22.20	23.70	25.20	26.70	28.00
13	9.385	10.72	12.58	14.34	16.03	17.67	19.25	20.80	22.30	23.80	25.30	26.70	28.10
14	9.000	10.76	12.64	14.40	16.09	17.72	19.31	20.86	22.40	23.90	25.30	26.80	28.20
15	8.933	10.80	12.68	14.44	16.14	17.78	19.35	20.90	22.40	23.90	25.40	26.80	28.20
16	8.79	10.84	12.72	14.48	16.18	17.81	19.40	20.90	22.50	24.00	25.40	26.90	28.30
17	8.81	10.87	12.74	14.52	16.22	17.85	19.50	21.00	22.50	24.00	25.40	26.90	28.30
18	8.84	10.90	12.78	14.56	16.25	17.87	19.50	21.10	22.60	24.10	25.50	26.90	28.30
19	8.86	10.92	12.81	14.58	16.27	17.90	19.50	21.10	22.60	24.10	25.50	27.00	28.40
20	8.87	10.94	12.83	14.60	16.30	18.00	19.50	21.10	22.60	24.10	25.50	27.00	28.40
∞	9.21	11.35	13.28	15.09	16.81	18.48	20.09	21.67	23.21	24.73	26.22	27.69	29.14

附表 9　Page 二因素分组检验统计量 L 的临界值表

（本表列出满足 $P\{L>l_\alpha\}\leqslant\alpha$ 的临界值 l_α.）

k	3			4		
		α			α	
n	0.001	0.01	0.05	0.001	0.01	0.05
2			28		60	58
3		42	41	89	87	84
4	56	55	54	117	114	111
5	70	68	66	145	141	137
6	83	81	79	172	167	163
7	96	93	91	198	193	189
8	109	106	104	225	220	214
9	121	119	116	252	246	240
10	134	131	128	278	272	266
11	147	144	141	305	298	292
12	160	156	153	331	324	317
13	172	169	165			
14	185	181	178			
15	197	194	190			
16	210	206	202			
17	223	218	215			
18	235	231	227			
19	248	243	239			
20	260	256	251			

k	5			6		
		α			α	
n	0.001	0.01	0.05	0.001	0.01	0.05
2	109	106	103	178	173	166
3	160	155	150	260	252	244
4	210	204	197	341	331	321
5	259	251	244	420	409	397
6	307	299	291	499	486	474
7	355	346	338	577	563	550
8	403	393	384	655	640	625
9	451	441	431	733	717	701
10	499	487	477	811	793	777
11	546	534	523	888	869	852
12	593	581	570	965	946	928

k	7			8		
		α			α	
n	0.001	0.01	0.05	0.001	0.01	0.05
2	269	261	252	388	376	362
3	394	382	370	567	549	532
4	516	501	487	743	722	701
5	637	620	603	917	893	869
6	757	737	719	1090	1063	1037
7	876	855	835	1262	1232	1204
8	994	972	950	1433	1401	1371
9	1113	1088	1065	1603	1569	1537
10	1230	1205	1180	1773	1736	1703
11	1348	1321	1295	1943	1905	1868
12	1465	1437	1410	2112	2072	2035

附表 10 Spearman 秩相关系数统计量 S 的临界值表

(本表列出满足 $P\{S\geqslant s_\alpha\}\leqslant\alpha(1)$(单边检验)

或 $P\{|S|\geqslant s_\alpha\}\leqslant\alpha(2)$(双边检验)的临界值 s_α.)

$\alpha(2)$	0.20	0.10	0.05	$\alpha(2)$	0.20	0.10	0.05
$\alpha(1)$	0.10	0.05	0.025	$\alpha(1)$	0.10	0.05	0.025
n				n			
4	1.000	1.000		29	0.245	0.312	0.368
5	0.800	0.900	1.000	30	0.240	0.306	0.362
6	0.657	0.829	0.886	31	0.236	0.301	0.356
7	0.571	0.714	0.786	32	0.232	0.296	0.350
8	0.524	0.643	0.738	33	0.229	0.291	0.345
9	0.483	0.600	0.700	34	0.225	0.287	0.340
10	0.455	0.564	0.648	35	0.222	0.283	0.335
11	0.427	0.536	0.618	36	0.219	0.279	0.330
12	0.406	0.503	0.587	37	0.216	0.275	0.325
13	0.385	0.484	0.560	38	0.212	0.271	0.321
14	0.367	0.464	0.538	39	0.210	0.267	0.317
15	0.354	0.446	0.521	40	0.207	0.264	0.313
16	0.341	0.429	0.503	41	0.204	0.261	0.309
17	0.328	0.414	0.485	42	0.202	0.257	0.305
18	0.317	0.401	0.472	43	0.199	0.254	0.301
19	0.309	0.391	0.460	44	0.197	0.251	0.298
20	0.299	0.380	0.447	45	0.194	0.248	0.294
21	0.292	0.370	0.435	46	0.192	0.246	0.291
22	0.284	0.361	0.425	47	0.190	0.243	0.288
23	0.278	0.353	0.415	48	0.188	0.240	0.285
24	0.271	0.344	0.406	49	0.186	0.238	0.282
25	0.265	0.337	0.398	50	0.184	0.235	0.279
26	0.259	0.331	0.390	51	0.182	0.233	0.276
27	0.255	0.324	0.382	52	0.180	0.231	0.274
28	0.250	0.317	0.375	53	0.179	0.228	0.271

续表 10

α(2)	0.20	0.10	0.05	α(2)	0.20	0.10	0.05
α(1)	0.10	0.05	0.025	α(1)	0.10	0.05	0.025
n				n			
54	0.177	0.226	0.268	78	0.147	0.188	0.223
55	0.175	0.224	0.266	79	0.146	0.186	0.221
56	0.174	0.222	0.264	80	0.145	0.185	0.220
57	0.172	0.220	0.261	81	0.144	0.184	0.219
58	0.171	0.218	0.259	82	0.143	0.183	0.217
59	0.169	0.216	0.257	83	0.142	0.182	0.216
60	0.168	0.214	0.255	84	0.141	0.181	0.215
61	0.166	0.213	0.252	85	0.140	0.180	0.213
62	0.165	0.211	0.250	86	0.139	0.179	0.212
63	0.163	0.209	0.248	87	0.139	0.177	0.211
64	0.162	0.207	0.246	88	0.138	0.176	0.210
65	0.161	0.206	0.244	89	0.137	0.175	0.209
66	0.160	0.204	0.243	90	0.136	0.174	0.207
67	0.158	0.203	0.241	91	0.135	0.173	0.206
68	0.157	0.201	0.239	92	0.135	0.173	0.205
69	0.156	0.200	0.237	93	0.134	0.172	0.204
70	0.155	0.198	0.235	94	0.133	0.171	0.203
71	0.154	0.197	0.234	95	0.133	0.170	0.202
72	0.153	0.195	0.232	96	0.132	0.169	0.201
73	0.152	0.194	0.230	97	0.131	0.168	0.200
74	0.151	0.193	0.229	98	0.130	0.167	0.199
75	0.150	0.191	0.227	99	0.130	0.166	0.198
76	0.149	0.190	0.226	100	0.129	0.165	0.197
77	0.148	0.189	0.224				

附表 11 Kendall 相关性检验统计量 K 的临界值表

（本表列出满足 $P\{K \geqslant k_\alpha\} \leqslant \alpha$ 的临界值 k_α.）

n	α		
	0.025	0.05	0.10
5	1.000	.800	.800
6	.867	.733	.600
7	.714	.619	.524
8	.643	.571	.429
9	.556	.500	.389
10	.511	.467	.378
11	.491	.418	.345
12	.455	.394	.303
13	.436	.359	.308
14	.407	.363	.275
15	.390	.333	.276
16	.383	.317	.250
17	.368	.309	.250
18	.346	.294	.242
19	.333	.287	.228
20	.326	.274	.221
21	.314	.267	.210
22	.307	.264	.203
23	.296	.257	.202
24	.290	.246	.196
25	.287	.240	.193
26	.280	.237	.188
27	.271	.231	.179
28	.265	.228	.180
29	.261	.222	.172
30	.255	.218	.172
31	.252	.213	.166
32	.246	.210	.165
33	.242	.205	.163
34	.237	.201	.159
35	.234	.197	.156
36	.232	.194	.152
37	.228	.192	.150
28	.223	.189	.149
39	.220	.188	.147
40	.218	.185	.144

附表 12　柯尔莫哥洛夫统计量

$$Q(\lambda)=\lim_{n\to\infty}P(D_n<\lambda/\sqrt{n})$$

λ	.00	.01	.02	.03	.04
0.2	0.000000	0.000000	0.000000	0.000000	0.000000
0.3	.000009	.000021	.000046	.000091	.000171
0.4	.002808	.003972	.005476	.007377	.009730
0.5	.036055	.042814	.050306	.058534	.067497
0.6	.135718	.149229	.163225	.177153	.192677
0.7	.288765	.305471	.322265	.339113	.355981
0.8	.455857	.472041	.488030	.503808	.519366
0.9	.607270	.620928	.634286	.647338	.666082
1.0	.730000	.740566	.750826	.760780	.770434
1.1	.822282	.829950	.837356	.844502	.851394
1.2	.887750	.893030	.898104	.902972	.907648
1.3	.931908	.935370	.938682	.941848	.944872
1.4	.960318	.962486	.964552	.966516	.968382
1.5	.977782	.979080	.980310	.981476	.982578
1.6	.988048	.988791	.989492	.990154	.990777
1.7	.993823	.994230	.994612	.994972	.995309
1.8	.996932	.997146	.997346	.997533	.997707
1.9	.998536	.998644	.998744	.998837	.998924
2.0	.999329	.999380	.999428	.999474	.999516
2.1	.999705	.999728	.999750	.999770	.999790
2.2	.999874	.999886	.999896	.999904	.999912
2.3	.999949	.999954	.999958	.999962	.999965
2.4	.999980	.999982	.999984	.999986	.999987

D_n 的极限分布表

$$=1-2\sum_{i=1}^{\infty}(-1)^{i+1}e^{-2i^2\lambda^2}$$

.05	.06	.07	.08	.09	λ
0.000000	0.000000	0.000000	0.000001	0.000004	0.2
.000303	.000511	.000826	.001285	.001929	0.3
.012590	.016005	.020022	.024682	.030017	0.4
.077183	.087577	.098656	.110395	.122760	0.5
.207987	.223637	.239582	.255780	.272189	0.6
.372833	.389640	.406372	.423002	.439505	0.7
.534682	.549744	.564546	.579070	.593316	0.8
.672516	.684630	.696444	.707940	.719126	0.9
.779794	.788860	.797636	.806128	.814342	1.0
.858038	.864442	.870612	.876548	.882258	1.1
.912132	.916432	.920556	.924505	.928288	1.2
.947756	.950512	.953142	.955650	.958040	1.3
.970158	.971846	.973448	.974970	.976412	1.4
.983622	.984610	.985544	.986426	.987260	1.5
.991364	.991917	.992438	.992928	.993389	1.6
.995625	.995922	.996200	.996460	.996704	1.7
.997870	.998023	.998145	.998297	.998421	1.8
.999004	.999079	.999149	.999213	.999273	1.9
.999552	.999588	.999620	.999650	.999680	2.0
.999806	.999822	.999838	.999852	.999864	2.1
.999920	.999926	.999934	.999940	.999944	2.2
.999968	.999970	.999973	.999976	.999978	2.3
.999988	.999988	.999990	.999991	.999992	2.4

附表 13 柯尔莫哥洛夫检验统计量 D_n 的临界值($d_{n\alpha}$)表

$P(D_n > d_{n\alpha}) = \alpha$

n \ α	0.20	0.10	0.05	0.02	0.01	n \ α	0.20	0.10	0.05	0.02	0.01
1	0.90000	0.95000	0.97500	0.99000	0.99500	31	0.18732	0.21412	0.23788	0.26596	0.28530
2	.68377	.77639	.84189	.90000	.92929	32	.18445	.21085	.23424	.26189	.28094
3	.56481	.63604	.70760	.78456	.82900	33	.18171	.20771	.23076	.25801	.28677
4	.49265	.56522	.62394	.68887	.73424	34	.17909	.20472	.22743	.25429	.27279
5	.44698	.50945	.56328	.62718	.66853	35	.17659	.20185	.22425	.25073	.26897
6	.41037	.46799	.51926	.57741	.61661	36	.17418	.19910	.22119	.24732	.26532
7	.38148	.43607	.48342	.53844	.57581	37	.17188	.19646	.21826	.24404	.26180
8	.35831	.40962	.45427	.50654	.54179	38	.16966	.19392	.21544	.24089	.25843
9	.33910	.38746	.43001	.47960	.51332	39	.16753	.19148	.21273	.23786	.25518
10	.32260	.36866	.40925	.45662	.48893	40	.16547	.18913	.21012	.23494	.25205
11	.30829	.35242	.39122	.43670	.46770	41	.16349	.18687	.20760	.23213	.24904
12	.29577	.33815	.37543	.41918	.44905	42	.16158	.18468	.20517	.22941	.24613
13	.28470	.32549	.36143	.40362	.43247	43	.15974	.18257	.20283	.22679	.24332
14	.27481	.31417	.34890	.38970	.41762	44	.15796	.18053	.20056	.22426	.24060
15	.26588	.30397	.33760	.37713	.40420	45	.15623	.17856	.19837	.22181	.23798
16	.25778	.29472	.32733	.36571	.39201	46	.15457	.17665	.19625	.21944	.23544
17	.25039	.28627	.31796	.35528	.38086	47	.15295	.17481	.19420	.21715	.23298
18	.24360	.27851	.30936	.34569	.37062	48	.15139	.17302	.19221	.21493	.23059
19	.23735	.27136	.30143	.33685	.36117	49	.14987	.17128	.19028	.21277	.22828
20	.23156	.26473	.29408	.32866	.35241	50	.14840	.16959	.18841	.21068	.22604
21	.22617	.25858	.28724	.32104	.34427	55	.14164	.16186	.17981	.20107	.21574
22	.22115	.25283	.28087	.31394	.33666	60	.13573	.15511	.17231	.19267	.20673
23	.21645	.24746	.27490	.30728	.32954	65	.13052	.14913	.16567	.18525	.19877
24	.21205	.24242	.26931	.30104	.32286	70	.12586	.14381	.15975	.17863	.19167
25	.20790	.23768	.26404	.29516	.31657	75	.12167	.13901	.15442	.17268	.18528
26	.20399	.23320	.25907	.28962	.31064	80	.11787	.13467	.14960	.16728	.17949
27	.20030	.22898	.25438	.28438	.30502	85	.11442	.13072	.14520	.16236	.17421
28	.19680	.22497	.24993	.27942	.29971	90	.11125	.12709	.14117	.15786	.16938
29	.19348	.22117	.24571	.27471	.29466	95	.10833	.12375	.13746	.15371	.16493
30	.19032	.21756	.24170	.27023	.28987	100	.10563	.12067	.13403	.14987	.16081

参 考 文 献

[1] R. H. Randles & D. A. Wolfe, *Introduction to the theory of nonparametric statistics*. Wiley & Sons, New York, 1979.

[2] P. J. Bickel & K. A. Doksum, Mathematical Statistics, Holden-Day, San Francisco, 1977.

[3] S. S. Wilks, *Mathematical Statistics*, Wiley & Sons, New York, 1962.

[4] 陈希孺,数理统计引论,科学出版社,北京,1981.

[5] 中国科学院数学研究所概率统计室,常用数理统计表,科学出版社,北京,1974.

[6] 陈家鼎,孙山泽,李东风,数理统计讲义,高等教育出版社,北京,1993.

[7] J. Hájek & Z. Šidák, *Theory of rank tests*, Academic Press, New York, 1967.

参 考 文 献

[1] R. H. Randles & D. A. Wolfe, *Introduction to the theory of nonparametric statistics*. Wiley & Sons, New York, 1979.

[2] P. J. Bickel & K. A. Doksum, Mathematical Statistics, Holden-Day, San Francisco, 1977.

[3] S. S. Wilks, *Mathematical Statistics*, Wiley & Sons, New York, 1962.

[4] 陈希孺,数理统计引论,科学出版社,北京,1981.

[5] 中国科学院数学研究所概率统计室,常用数理统计表,科学出版社,北京,1974.

[6] 陈家鼎,孙山泽,李东风,数理统计讲义,高等教育出版社,北京,1993.

[7] J. Hájek & Z. Šidák, *Theory of rank tests*, Academic Press, New York, 1967.

参 考 文 献

[1] R. H. Randles & D. A. Wolfe, *Introduction to the theory of nonparametric statistics*. Wiley & Sons, New York, 1979.

[2] P. J. Bickel & K. A. Doksum, Mathematical Statistics, Holden-Day, San Francisco, 1977.

[3] S. S. Wilks, *Mathematical Statistics*, Wiley & Sons, New York, 1962.

[4] 陈希孺,数理统计引论,科学出版社,北京,1981.

[5] 中国科学院数学研究所概率统计室,常用数理统计表,科学出版社,北京,1974.

[6] 陈家鼎,孙山泽,李东风,数理统计讲义,高等教育出版社,北京,1993.

[7] J. Hájek & Z. Šidák, *Theory of rank tests*, Academic Press, New York, 1967.

参考文献

[1] R. H. Randles & D. A. Wolfe, *Introduction to the theory of nonparametric statistics*. Wiley & Sons, New York, 1979.

[2] P. J. Bickel & K. A. Doksum, Mathematical Statistics, Holden-Day, San Francisco, 1977.

[3] S. S. Wilks, *Mathematical Statistics*, Wiley & Sons, New York, 1962.

[4] 陈希孺,数理统计引论,科学出版社,北京,1981.

[5] 中国科学院数学研究所概率统计室,常用数理统计表,科学出版社,北京,1974.

[6] 陈家鼎,孙山泽,李东风,数理统计讲义,高等教育出版社,北京,1993.

[7] J. Hájek & Z. Šidák, *Theory of rank tests*, Academic Press, New York, 1967.

参 考 文 献

[1] R. H. Randles & D. A. Wolfe, *Introduction to the theory of nonparametric statistics*. Wiley & Sons, New York, 1979.

[2] P. J. Bickel & K. A. Doksum, Mathematical Statistics, Holden-Day, San Francisco, 1977.

[3] S. S. Wilks, *Mathematical Statistics*, Wiley & Sons, New York, 1962.

[4] 陈希孺,数理统计引论,科学出版社,北京,1981.

[5] 中国科学院数学研究所概率统计室,常用数理统计表,科学出版社,北京,1974.

[6] 陈家鼎,孙山泽,李东风,数理统计讲义,高等教育出版社,北京,1993.

[7] J. Hájek & Z. Šidák, *Theory of rank tests*, Academic Press, New York, 1967.